(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
24 February 2005 (24.02.2005)

PCT

(10) International Publication Number
WO 2005/016230 A2

(51) International Patent Classification[7]: A61K

(21) International Application Number: PCT/US2003/017979

(22) International Filing Date: 9 June 2003 (09.06.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/387,034 7 June 2002 (07.06.2002) US

(71) Applicant (*for all designated States except US*): **PRESIDENT AND FELLOWS OF HARVARD COLLEGE** [US/US]; 17 Quincy Street, Cambridge, MA 02138-3876 (US).

(72) Inventor; and
(75) Inventor/Applicant (*for US only*): **LABAER, Joshua** [US/US]; 22 Moraine Street, Jamaica Plain, MA 02130-4316 (US).

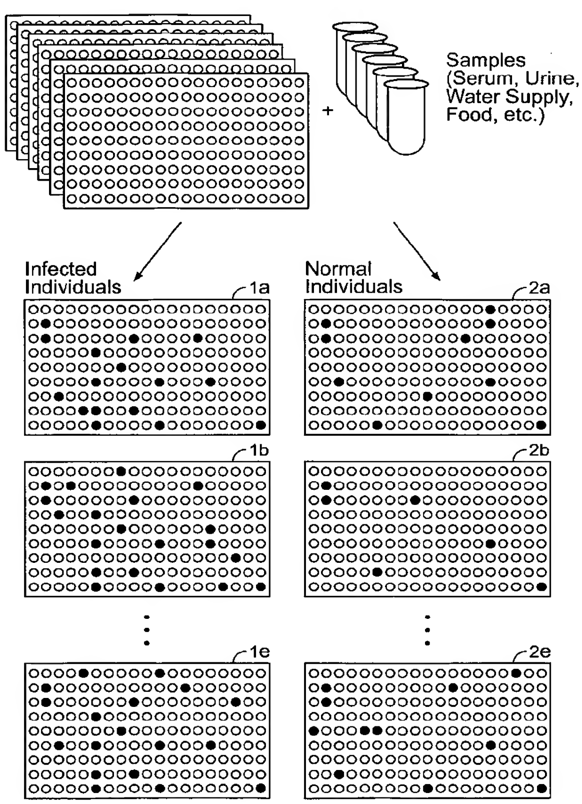
(74) Agent: **MYERS, Louis**; Fish & Richardson P.C., 225 Franklin Street, Boston, MA 02110-2804 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO,

[Continued on next page]

(54) Title: EVALUATING PROTEIN SIGNATURES



(57) Abstract: Test arrays of capture probes are used to identify characteristic information about a sample. For example, the methods can be used to identify the presence of a cancer cell or a pathogen in a sample from a subject, or the presence of a target molecule in an environmental sample.

WO 2005/016230 A2

SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

EVALUATING PROTEIN SIGNATURES

Cross-Reference to Related Applications

This application claims priority to U.S. provisional application 60/387,034, filed June 7, 2002, which is herein incorporated by reference in its entirety.

Government Support

Aspects of the invention were made, in part, with government support under grant R01 DK61906 awarded by the National Institutes of Health. The government may have certain rights in the invention.

Background

Pathogens (and diseased cells) have macromolecules that are unique, i.e., are not normally found in a biological specimen from a subject not having the pathogen or disease. In addition, infected subjects and affected cells produce response proteins that also are signals of the disease state.

The use of marker proteins has been demonstrated in a number of cases of disease or infection, for example: HBSAg (a protein indicating the presence of active type B hepatitis), p24 (used to detect the presence of HIV), CA-125 (used to detect ovarian cancer and some lung cancers), CMV Antigen (Cytomegalovirus detection), Cryptococcal Antigen (detection of cryptococcal infection), Rheumatoid Factor (rheumatoid arthritis), etc. However, the number of situations in which marker proteins have been exploited has been surprisingly limited, given the theoretical likelihood identifying such proteins in most pathogens and diseases.

Reasons why this approach has not been applied more broadly include difficulty of identifying proteins or macromolecules that are sufficiently specific to indicate a particular pathogen or disease, and that the presence or absence of a single protein may not be sufficiently predictive to make a definitive diagnosis. There are thousands of proteins from which to choose, and no routine methods are available for identifying the best candidates. For example, CA-125 is unusual in healthy women, and is often elevated in ovarian cancer. However, the same antigen is also elevated in some lung cancers.

A routine method for identifying marker proteins for a disease, to determine the presence of a pathogen, or specifically diagnose a pathological condition, remains a basic need of diagnostic medicine.

Summary

In one aspect, the invention features a method that includes: preparing a binding pattern signature for the target by contacting a first sample of a test array with a positive control for the target. The test array has a plurality of proteins affixed to a substrate at replicable locations, wherein identification of compositions of the proteins need not be known. A second sample of the array is contacted with a negative control not containing the target. Conditions can be used such that components in the positive control detectably bind to locations of the first array and produce a target pattern, and components in the negative control detectably bind to locations of the second array and produce a control pattern. The signature for the target biological material has locations present in the target pattern and absent from the control pattern. Optionally, a third sample of the array is contacted with the specimen, to obtain a specimen pattern of locations and the specimen pattern is compared with the target signature, so that presence of the target in the specimen is indicated by the presence of the target signature in the specimen pattern. The target material can be of biological origin, such as a macromolecule, such as all or part of a protein, carbohydrate, lipid, or a monomer thereof, or a molecule having components of more than one of these, such as a lipoprotein. The method can be used, e.g., for evaluating a target, e.g., a sample or a target material in a specimen.

In an embodiment of the method, the proteins affixed to the substrate are from a mammal, for example, the proteins affixed to the substrate are from a human, for example, the proteins affixed to the substrate are from a cancer cell. "Proteins" on the array include portions of proteins, such as peptides and oligopeptides, which terms are used interchangeably and have the same meaning.

The cancer cell is, for example, a primary or metastatic cancer of lung, skin, leukemia, lymphoma, brain, breast, prostate, bowel, esophagus, liver, pancreas, and head or neck cancers. In a different embodiment, the proteins affixed to the substrate are from a pathogen. For example, the pathogen is a virus, a bacterium, a fungus, or a protozoan. The target can be a prion. The protein on the array from a bacterium can be any species of a genus *Actinobacillus*, *Bacillus*, *Borrelia*, *Brucella*, *Chlamydia*,

Clostridium, *Coxiella*, *Enterococcus*, *Escherichia*, *Francisella*, *Hemophilus*, *Legionella*, *Mycobacterium*, *Neisseria*, *Pasteurella*, *Pseudomonas*, *Salmonella*, *Shigella*, *Staphylococcus*, *Streptococcus*, *Treponema*, or *Yersinia*. For example, the *Bacillus* is *B. anthracis;* the *Escherichia* is *E. coli* O157:H7, *the Mycobacterium* is *M. tuberculosis;* and the *Borrelia* is *B. burgdorferi*. Further, the proteins can be from a spore, for example, of *Bacillus anthracis*.

In an alternative embodiment virus is influenza, human immunodeficiency, Venezuelan equine encephalitis, West Nile, smallpox, rhinovirus, Ebola, Rift Valley fever, Lassa fever, measles, mumps, Marburg, yellow fever, herpes, hantavirus, hepatitis A, hepatitis B, hepatitis C, rotavirus, parvovirus, rabies, respiratory syncytial, rubella, Epstein Barr, Newcastle disease, hoof and mouth, tobacco mosaic, Glycine mosaic comovirus, or wheat American striate. Further, the protein on an array from a fungus is a species from the group of genera *Aspergillus*, *Candida*, *Phytophthora*, *Puccinia*, *Lichen*, and *Trichophyton*. For example, the *Aspergillus* is *A. flavus*. The target material can be a bacterial or fungal toxin. The protein on an array from a protozoan is *Plasmodium*, *Leishmania*, *Entamoeba*, *Enterocytozoan*, *Cryptosporidium*, and *Giardia*. Alternatively, the proteins affixed to the substrate are random proteins.

The specimen can be a biological fluid sample, for example, urine, saliva, lacrymal secretions, nasal discharge, blood, serum, plasma, lymph, perspiration, amniotic fluid, cerebrospinal fluid, ascites fluid, semen, vaginal secretions, feces, or cell extract. Alternatively, the specimen is an environmental sample, for example, the environmental sample is a soil suspension, air infusion, pond water, lake water, river water, ocean water, sewage, industrial effluent, food, beverages, consumable goods, packaged goods, mail, baggage, a fluid extract of a rubbing or an instrument or object, or any physical sample in any phase (e.g., liquid, solid, or vapor).

The preparing step in certain embodiments is re-iterated to obtain a statistically significant number of signature binding patterns for the target material. The method can be re-iterated for one or more additional target materials, for example, the signature binding pattern for the additional biological material is obtained using a fourth sample of the replicable test array. The preparing step can in certain embodiments be preparing a binding signature for a target material which is a component of a novel organism. Iteration can be performed in parallel or sequentially.

Further, the biological fluid can be obtained from a patient with an acute medical condition, for example, the acute medical condition is a cardiac condition, such as myocardial infarction or stroke. Alternatively, the biological fluid can be obtained from a patient with an autoimmune disease, such as multiple sclerosis, myasthenia gravis, Hashimoto's disease, systemic lupus erythematosis, uveitis, Guillain-Barre' syndrome, Grave's disease, idiopathic myxedema, autoimmune oophoritis, chronic immune thrombocytopenic purpura, colitis, diabetes, psoriasis, pemphigus vulgaris, and rheumatoid arthritis. Further, the sample can be obtained from a patient having an inflammatory condition, for example, asthma, allergy, and inflammatory bowel syndrome.

In another embodiment, contacting the second sample of the array with the negative control, the signature for the target biological material additionally comprises at least one location present in the control pattern and absent from the target pattern. An embodiment of the inventions is thus a universal test array for detecting an unwanted cell, disease, or organism in a biological specimen from a mammal by a method described herein (e.g., the above method), wherein the sample of the test array comprises a plurality of proteins of the mammal.

The invention also provides a kit for detecting a pathogen, a kit for detecting a cancer cell, a kit for detecting an acute medical condition, and a kit for detecting an autoimmune disease, e.g., a kit provided by a method described herein. In one embodiment, the kit includes: (1) an array comprising a plurality of addresses, wherein each address of the plurality comprises a handle and (2) a vector nucleic acid comprising (i) a promoter; (ii) an entry site; and (iii) a tag encoding sequence, wherein the tag can be attached to the handle. The kit can further include software and/or a database, e.g., in computer memory or a computer readable medium (e.g., a CD-ROM, a magnetic disc, flash memory. Each record of the database can include a field for the polypeptide (e.g., a randomized polypeptide) encoded by the nucleic acid sequence and a descriptor or reference for the physical location of the encoding nucleic acid sequence in the kit, e.g., location in a microtitre plate. Optionally, the record also includes a field representing a result (e.g., a qualitative or quantitative result) of detecting the polypeptide encoded by the nucleic acid sequence. The database can include a record for each address of the plurality present on the array. The records can be clustered or have a reference to other records (e.g., including hierarchical groupings) based on the

result. The software can contain computer readable code to configure a computer-controlled robotic apparatus to manipulate nucleic acids encoding test amino acid sequences and vector nucleic acids in order to insert the encoding nucleic acids into the vector nucleic acids and further to manipulate the insertion products onto addresses of the array. The kit can also include instructions for use of the array or a link or indication of a network resource (e.g., a web site) having instructions for use of the array or the above database of records describing the addresses of the array.

A method of providing an array can include: providing the aforementioned kit, and a plurality of nucleic acid sequences, each encoding a unique test amino acid sequence and an excision site. The method further includes removing each of the plurality of nucleic acid sequence from the excision site and inserting it into the entry site of the vector nucleic acid to thereby generate a test nucleic acid sequence encoding a test polypeptide comprising the test amino acid sequence and the tag; and disposing each of the plurality of test nucleic acid sequences at an address of the array.

In one embodiment, the proteins on the array are randomized or include a randomized segment of at least 10 amino acid in which at least four, five, eight, nine or ten positions are randomized. Randomization can be generated, e.g., using degenerate oligonucleotides or a random number generator (e.g., on a computer). In one embodiment, the randomized segment is between 10-50, 10-20, or 10-200 amino acids in length. Longer segments can also be used. The degree of randomization can vary at a given position and by the number of positions (e.g., between 10-100, 30-100, 60-100, 80-100 or 70-90% of the positions. Randomization can included biased compositions of starting material. In one embodiment, the randomized segment is within a domain of a folded protein, e.g., a binding loop of an extracellular protein or domain thereof, e.g., a CDR of an immunoglobulin domain.

In another aspect, the invention features a method of providing an interaction profile. The method includes providing an array of capture probes, contacting a sample to the array, and identifying probes to which the sample interacts (e.g., to which one or more molecules in the sample interacts), thus providing an interaction profile. The array includes a plurality of capture probes. Each capture probe is positionally distinguishable from the other probes. In one embodiment, each probe includes a unique region. In another embodiment, each probe includes a randomized region.

In one embodiment, the interaction of the compound with the probe results in a covalent modification of the probe, e.g., a covalent bond of the probe can be broken or formed. In a preferred embodiment, the interaction of the compound with the capture probe is a binding interaction wherein neither the compound nor the probe has a covalent bond broken or formed.

In a preferred embodiment, the interaction profile is a list of objects, each object representing a unique capture probe, and having an associated parameter, e.g., a numerical value. The list can contain two, three, four, five, six, seven, eight, nine, ten, 15, 20, 50, 100, 1000 or more objects. In a preferred embodiment, each unique capture probe is represented by an object. In this embodiment, the list includes as many objects as unique capture probes. In another embodiment, the list includes the capture probes which interact with the compound. Thus, the list can contain only those capture probes for which an interaction was detected, or only those capture probes for which an interaction met a predetermined condition. Such a list has fewer objects as members than the number of unique capture probes. In a preferred embodiment, the interaction profile is stored in computer memory, such as random access memory or flash memory, or on computer readable media, such as magnetic (e.g., a diskette, removable hard drive, or internal hard drive) or optical media (e.g., a compact disk (CD), DVD, or holographic media). A profile stored in this manner can be on a personal computer, server, e.g., a network server, or mainframe, and can be accessed from another device across a network, e.g., an intranet or internet. In another embodiment, the interaction profile is printed on to a media such as a plastic, a paper or a label, e.g., as a bar code or variation thereof.

The parameter associated with each object of an interaction profile can be obtained from a quantitative observation, or a qualitative observation, preferably a quantitative observation. In one embodiment, the associated parameter is a function of the amount of interaction between the compound and the probe. For example, the amount of interaction can be the amount of binding, the amount of probe modification, or affinity. In a preferred embodiment, the associated parameter is a function of the amount of binding between the compound and the probe. The parameter can be a function of the amount of a quantitative observation such as a fluorescent signal, a radioactive signal, or a phosphorescent signal of a contacted capture probe. The parameter can be provided by an instrument, e.g., a CCD camera. In one embodiment,

the parameter is a function of the surface plasmon resonance at the site of a contacted capture probe. In a preferred embodiment, the associated parameter are adjusted for a background signal. In another embodiment, the associated parameter is a function of moles of bound compound. In yet another embodiment, the associated parameter is an affinity, relative affinity, apparent affinity, association constant, dissociation constant, logarithm of an affinity, or free energy for binding, of the compound for the capture probe. In a preferred embodiment, the associated parameter in the list are differ. In other words, the list contains more than one object, and i.e., the associated parameter of the objects in the list are not all the same. Where the associated parameters are values, the values can provide a range. The values can be distributed in the range. In some embodiments, the values can approximate a Poisson distribution. The list can contain objects whose associated values are zero, or null. The list can contain objects whose associated values are positive or negative. In one embodiment, the list does not contain any objects whose associated values are zero or null.

In an embodiment, interaction profiles are provided for a sample using varying amounts of the sample, i.e. an interaction profile is provided for a sample at a first concentration, at a second concentration, etc. In another preferred embodiment, interaction profiles are provided for a sample for interaction with varying concentrations of capture probes. For example, an array can have more than one unit, the compositions of the units being identical, but the first unit having the probes at a first concentration, and the second unit having the probes at a second concentration, etc. In yet another preferred embodiment, interaction profiles are provide for a sample for various intervals after contacting the sample to the array. For example, a first profile can be provided after a first interval of time has elapsed after application, and a second profile can be provided after a second interval, etc.

The capture probes contain a unique region. In one embodiment, the unique region is an interaction site, an interaction site variant, or putative interaction site. In another embodiment, the unique region is random.

The array of the present invention can contain at least two, four, preferably 16, 64, 96, 128, 384, 1536, or more unique capture probes. In one embodiment, the array is a solid silica support. The plurality of nucleic acid probes can be stably attached to the support by a covalent bond. For example, a probe can be stably attached with a silane compound reactive with primary amines, or acrylamide moieties in the oligonucleotide

or PCR productions, or an amino silane or polylysine or other polymer capable of UV crosslinking to single-stranded or double-stranded DNA. In a preferred embodiment, the array is a glass slide. The slide can be treated with an activating agent, e.g., 1,4-diphenylene-diisothiocyanate (PDC).

In one embodiment, the capture probes are small organic chemicals, e.g., compounds with a molecular weight of 10,000, 5,000, 3,000, or 1,000 Daltons or less. The chemicals can be produced by combinatorial synthesis.

In a preferred embodiment, the capture probes on the array can be biological polymers such as nucleic acids, polypeptides, complex sugars, and combinations thereof. For example, a polypeptide can be covalently linked to a DNA or RNA. In one embodiment, the capture probes are polypeptides. The polypeptides can contain 2, 10, 20, 30, 50, or 100 or more amino acids. In a preferred embodiment, the polypeptides are antibodies. In a preferred embodiment, the capture probes are nucleic acids. The capture probes can be a deoxyribonucleic acid (DNA), a ribonucleic acid (RNA), a peptide nucleic acid (PNA), or any combination thereof. In embodiments in which the monitored interaction between the sample and the capture probes is a binding interaction, the unique region is preferably a binding site.

As used herein, the term "database" refers to at least one table of information, containing at least one record. A record is a row in the table. A record can have one or more fields or attributes. For example, in a database of interaction profiles, a record can have fields describing the location of a capture probe on an array, the composition, e.g., nucleic acid sequence, of the capture probe at the location, and/or a value, e.g., a numerical value, which is a function of the extent of interaction of the capture probe with a sample.

Arrays of polypeptides can be generated by translation of nucleic acid sequences encoding the polypeptides at individual addresses on the array. This allows for the rapid and versatile development of a polypeptide microarray platform for analyzing and manipulating biological information.

In one aspect, the invention features an array including a substrate having a plurality of addresses. Each address of the plurality includes: (1) a nucleic acid (e.g., a DNA or an RNA) encoding a hybrid amino acid sequence which includes a test amino acid sequence and an affinity tag; and, optionally, (2) a binding agent that recognizes

the affinity tag. Optionally, each address of the plurality also includes one or both of (i) an RNA polymerase; and (ii) a translation effector.

In a preferred embodiment, each test amino acid sequence in the plurality of addresses is unique. For example, a test amino acid sequence can differ from all other test amino acid sequence of the plurality by 1, or more amino acid differences, (e.g., about 2, 3, 4, 5, 8, 16, 32, 64 or more differences; and, by way of example, has about 800, 256, 128, 64, or 32, 16, 8, 4, or fewer differences). In another preferred embodiment, the test amino acid sequence encoded by the nucleic acid at each address of the plurality is identical to all other test amino acid sequences in the plurality of addresses. In a preferred embodiment, the affinity tag encoded by the nucleic acid at each address of the plurality is the same, or substantially identical to all other affinity tags in the plurality of addresses. In another preferred embodiment, the nucleic acid at each address of the plurality encodes more than one affinity tag. In yet another preferred embodiment, the affinity tag encoded by the nucleic acid at an address of the plurality differs from at least one other affinity tag in the plurality of addresses.

In a preferred embodiment, the affinity tag is fused directly to the test amino acid sequence, e.g., directly amino-terminal, or directly carboxy-terminal. In another preferred embodiment, the affinity tag is separated from the test amino acid by one or more linker amino acids, e.g., 1, 2, 3, 4, 5, 6, 8, 10, 12, 20, 30 or more amino acids, preferably about 1 to 20, or about 3 to 12 amino acids. The linker amino acids can include a cleavage site, flexible amino acids (e.g., glycine, alanine, or serine, preferably glycine), and/or polar amino acids. The linker and affinity tag can be amino-terminal or carboxy-terminal to the test amino acid sequence.

The nucleic acid can be a RNA, or a DNA (e.g., a single-stranded DNA, or a double stranded DNA). In a preferred embodiment, the nucleic acid includes a plasmid DNA or a fragment thereof; an amplification product (e.g., a product generated by RCA, PCR, NASBA); or a synthetic DNA.

The nucleic acid can further include one or more of: a transcription promoter; a transcription regulatory sequence; a untranslated leader sequence; a sequence encoding a cleavage site; a recombination site; a 3' untranslated sequence; a transcriptional terminator; and an internal ribosome entry site. In one embodiment, the nucleic acid sequence includes a plurality of cistrons (also termed "open reading frames"), e.g., the sequence is dicistronic or polycistronic. In another embodiment, the nucleic acid also

includes a sequence encoding a reporter protein, e.g., a protein whose abundance can be quantitated and can provide an indication of the quantity of test polypeptide fixed to the plate. The reporter protein can be attached to the test polypeptide, e.g., covalently attached, e.g., attached as a translational fusion. The reporter protein can be an enzyme, e.g., β-galactosidase, chloramphenicol acetyl transferase, β-glucuronidase, and so forth. The reporter protein can produce or modulate light, e.g., a fluorescent protein (e.g., green fluorescent protein, variants thereof, red fluorescent protein, variants thereof, and the like), and luciferase.

The transcription promoter can be a prokaryotic promoter, a eukaryotic promoter, or a viral promoter. In a preferred embodiment, the promoter is the T7 RNA polymerase promoter. The regulatory components, e.g., the transcription promoter, can vary among nucleic acids at different addresses of the plurality. For example, different promoters can be used to vary the amount of polypeptide produced at different addresses.

In one embodiment, the nucleic acid also includes at least one site for recombination, e.g., homologous recombination or site-specific recombination, e.g., a lambda att site or variant thereof; a lox site; or a FLP site. In a preferred embodiment, the recombination site lacks stop codons in the reading frame of a nucleic acid encoding a test amino acid sequence. In another preferred embodiment, the recombination site includes a stop codon in the reading frame of a nucleic acid encoding a test amino acid sequence.

In another embodiment, the nucleic acid includes a sequence encoding a cleavage site, e.g., a protease site, e.g., a site cleaved by a site-specific protease (e.g., a thrombin site, an enterokinase site, a PreScission site, a factor Xa site, or a TEV site), or a chemical cleavage site (e.g., a methionine, preferably a unique methionine (cleavage by cyanogen bromide) or a proline (cleavage by formic acid)).

The nucleic acid can include a sequence encoding a second polypeptide tag in addition to the affinity tag. The second tag can be C-terminal to the test amino acid sequence and the affinity tag can be N-terminal to the test amino acid sequence; the second tag can be N-terminal to the test amino acid sequence, and the affinity tag can be C-terminal to the test amino acid sequence; the second tag and the affinity tag can be adjacent to one another, or separated by a linker sequence, both being N-terminal or C-terminal to the test amino acid sequence. In one embodiment, the second tag is an

additional affinity tag, e.g., the same or different from the first tag. In another embodiment, the second tag is a recognition tag. For example, the recognition tag can report the presence and/or amount of test polypeptide at an address. Preferably the recognition tag has a sequence other than the sequence of the affinity tag. In still another embodiment, a plurality of polypeptide tags (e.g., less than 3, 4, 5, about 10, or about 20 tags) are encoded in addition to the first affinity tag. Each polypeptide tag of the plurality can be the same as or different from the first affinity tag.

The nucleic acid sequence can further include an identifier sequence, e.g., a non-coding nucleic acid sequence, e.g., one that is synthetically inserted, and allows for uniquely identifying the nucleic acid sequence. The identifier sequence can be sufficient in length to uniquely identify each sequence in the plurality; e.g., it is about 5 to 500, 10 to 100, 10 to 50, or about 10 to 30 nucleotides in length. The identifier can be selected so that it is not complementary or identical to another identifier or any region of each nucleic acid sequence of the plurality on the array.

The test amino acid sequence can further include a protein splicing sequence or intein. The intein can be inserted in the middle of a test amino acid sequence. The intein can be a naturally-occurring intein or a mutated intein.

The nucleic acids encoding the test amino acid sequences can be obtained from a collection of full-length expressed genes (e.g., a repository of clones), a cDNA library, or a genomic library. The encoding nucleic acids can be nucleic acids (e.g., an mRNA or cDNA) expressed in a tissue, e.g., a normal or diseased tissue. The test polypeptides (i.e., test amino acid sequences) can be mutants or variants of a scaffold protein (e.g., an antibody, zinc-finger, polypeptide hormone etc.). In yet another embodiment, the test polypeptides are random amino acid sequences, patterned amino acids sequences, or designed amino acids sequences (e.g., sequence designed by manual, rational, or computer-aided approaches). The plurality of test amino acid sequences can include a plurality from a first source, and plurality from a second source. For example, the test amino acid sequences on half the addresses of an array are from a diseased tissue or a first species, whereas the sequences on the remaining half are from a normal tissue or a second species.

In a preferred embodiment, each address of the plurality further includes one or more second nucleic acids, e.g., a plurality of unique nucleic acids. Hence, the plurality in toto can encode a plurality of test sequences. For example, each address of the

plurality can encode a pool of test polypeptide sequences, e.g., a subset of a library or clone bank. A second array can be provided in which each address of the plurality of the second array includes a single or subset of members of the pool present at an address of the first array. The first and the second array can be used consecutively.

In other preferred embodiments, each address of the plurality further includes a second nucleic acid encoding a second amino acid sequence.

In one preferred embodiment, each address of the plurality includes a first test amino acid sequence that is common to all addresses of the plurality, and a second test amino acid sequence that is unique among all the addresses of the plurality. For example, the second test amino acid sequences can be query sequences whereas the first amino test amino acid sequence can be a target sequence. In another preferred embodiment, each address of the plurality includes a first test amino acid sequence that is unique among all the addresses of the plurality, and a second test amino acid sequence that is common to all addresses of the plurality. For example, the first test amino acid sequences can be query sequences whereas the second amino test amino acid sequence can be a target sequence. The second nucleic acid encoding the second test amino acid sequence can include a sequence encoding a recognition tag and/or an affinity tag.

At at least one address of the plurality, the first and second amino acid sequences can be such that they interact with one another. In one preferred embodiment, they are capable of binding to each other. The second test amino acid sequence is optionally fused to a detectable amino acid sequence, e.g., an epitope tag, an enzyme, a fluorescent protein (e.g., GFP, BFP, variants thereof). The second test amino acid sequence can be itself detectable (e.g., an antibody is available which specifically recognizes it). In another preferred embodiment, one is capable of modifying the other (e.g., making or breaking a bond, preferably a covalent bond, of the other). For example, the first amino acid sequence is kinase capable of phosphorylating the second amino acid sequence; the first is a methylase capable of methylating the second; the first is a ubiquitin ligase capable of ubiquitinating the second; the first is a protease capable of cleaving the second; and so forth.

These embodiments can be used to identify an interaction or to identify a compound that modulates, e.g., inhibits or enhances, an interaction.

The binding agent can be attached to the substrate. For example, the substrate can be derivatized and the binding agent covalent attached thereto. The binding agent can be attached via a bridging moiety, e.g., a specific binding pair. (e.g., the substrate contains a first member of a specific binding pair, and the binding agent is linked to the second member of the binding pair, the second member being attached to the substrate).

In yet another embodiment, an insoluble substrate (e.g., a bead or particle), is disposed at each address of the plurality, and the binding agent is attached to the insoluble substrate. The insoluble substrate can further contain information encoding its identity, e.g., a reference to the address on which it is disposed. The insoluble substrate can be tagged using a chemical tag, or an electronic tag (e.g., a transponder). The insoluble substrate can be disposed such that it can be removed for later analysis.

Also featured is a database, e.g., in computer memory or a computer readable medium. Each record of the database can include a field for the amino acid sequence encoded by the nucleic acid sequence and a descriptor or reference for the physical location of the nucleic acid sequence on the array. Optionally, the record also includes a field representing a result (e.g., a qualitative or quantitative result) of detecting the polypeptide encoded by the nucleic acid sequence. The database can include a record for each address of the plurality present on the array. The records can be clustered or have a reference to other records (e.g., including hierarchical groupings) based on the result.

In another aspect, the invention features an array including a substrate having a plurality of addresses. Each address of the plurality includes: (1) an RNA encoding a hybrid amino acid sequence comprising a test amino acid sequence and an affinity tag; and (2) a binding agent that recognizes the affinity tag. Optionally, each address of the plurality also includes one or both of (i) a transcription effector; and (ii) a translation effector. The array can include other features described herein.

In another aspect, the invention features a method of providing an array of polypeptides. The method includes: (1) providing or obtaining a substrate with a plurality of addresses, each address of the plurality including (i) a nucleic acid encoding an amino acid sequence comprising a test amino acid sequence and an affinity tag, and (ii) a binding agent that recognizes the affinity tag; (2) contacting each address of the plurality with a translation effector to thereby translate the hybrid amino acid sequence; and (3) maintaining the substrate under conditions permissive for the amino acid

sequence to bind the binding agent. The substrate can be contacted to a sample, e.g., as described here.

In one embodiment, the nucleic acid provided on the substrate is synthesized in situ, e.g., by light-directed chemistry. In another embodiment, each address of the plurality is provided with a nucleic acid, e.g., by pipetting, spotting, printing (e.g., with pins), piezoelectric delivery, or, e.g., other means of mechanical delivery. In a preferred embodiment, the provided nucleic acid is a template nucleic acid, and the method further includes amplifying the template, e.g., by PCR, NASBA, or RCA. The method can further include transcribing the nucleic acid to produce one or more RNA molecules encoding the test amino acid sequence.

The method can further include washing the substrate, e.g., after sufficient contact with a translation effector. The wash step can be repeated, e.g., one or more times, e.g., until a translation effector or translation effector component is removed. The wash step can remove unbound proteins. The stringency of the wash step can vary, e.g., the salt, pH, and buffer composition of the wash buffer can vary. For example, if the translated test polypeptide is covalently captured, or captured by an interaction resistant to chaotropes (e.g., binding of a 6-histidine motif to Ni^{2+}·NTA), the substrate can be washed with a chaotrope, (e.g., guanidinium hydrochloride, or urea). In a subsequent step, the chaotrope can itself be washed from the array, and the polypeptides renatured.

In one embodiment, the nucleic acid sequence also encodes a cleavage site, e.g., a protease site, e.g., between the test amino acid sequence and the affinity tag. The method can further include contacting an address of the array with a protease that specifically recognizes the site.

The method can further include contacting the substrate with a second substrate. For example, in an embodiment wherein the substrate is a gel, the gel can be contacted with a second gel, and the contents of one gel can be transferred to another (e.g., by diffusion or electrophoresis). The method can include disrupting the binding between the affinity tag and the binding agent or between the binding agent and the substrate prior to transfer.

The method can further include contacting the substrate with living cells, and detecting an address wherein a parameter of the cell is altered relative to another address.

In a preferred embodiment, each test amino acid sequence in the plurality of addresses is unique. For example, a test amino acid sequence can differ from all other test amino acid sequence of the plurality by 1, or more amino acid differences, (e.g., about 2, 3, 4, 5, 8, 16, 32, 64 or more differences; and, by way of example, has about 800, 256, 128, 64, or 32, 16, 8, 4, or fewer differences). In another preferred embodiment, the test amino acid sequence encoded by the nucleic acid at each address of the plurality is identical to all other test amino acid sequences in the plurality of addresses. In a preferred embodiment, the affinity tag encoded by the nucleic acid at each address of the plurality is the same, or substantially identical to all other affinity tags in the plurality of addresses. In another preferred embodiment, the nucleic acid at each address of the plurality encodes more than one affinity tag. In yet another preferred embodiment, the affinity tag encoded by the nucleic acid at an address of the plurality differs from at least one other affinity tag in the plurality of addresses.

In a preferred embodiment, the affinity tag is fused directly to the test amino acid sequence, e.g., directly amino-terminal, or directly carboxy-terminal. In another preferred embodiment, the affinity tag is separated from the test amino acid by one or more linker amino acids, e.g., 1, 2, 3, 4, 5, 6, 8, 10, 12, 20, 30 or more amino acids, preferably about 1 to 20, or about 3 to 12 amino acids. The linker amino acids can include a cleavage site, flexible amino acids (e.g., glycine, alanine, or serine, preferably glycine), and/or polar amino acids. The linker and affinity tag can be amino-terminal or carboxy-terminal to the test amino acid sequence.

The nucleic acid can further include one or more of: a transcription promoter; a transcription regulatory sequence; a untranslated leader sequence; a sequence encoding a cleavage site; a recombination site; a 3' untranslated sequence; a transcriptional terminator; and an internal ribosome entry site. In one embodiment, the nucleic acid sequence includes a plurality of cistrons (also termed "open reading frames"), e.g., the sequence is dicistronic or polycistronic. In another embodiment, the nucleic acid also includes a sequence encoding a reporter protein, e.g., a protein whose abundance can be quantitated and can provide an indication of the quantity of test polypeptide fixed to the plate. The reporter protein can be attached to the test polypeptide, e.g., covalently attached, e.g., attached as a translational fusion. The reporter protein can be an enzyme, e.g., β-galactosidase, chloramphenicol acetyl transferase, β-glucuronidase, and so forth. The reporter protein can produce or modulate light, e.g., a fluorescent protein

(e.g., green fluorescent protein, variants thereof, red fluorescent protein, variants thereof, and the like), and luciferase.

The transcription promoter can be a prokaryotic promoter, a eukaryotic promoter, or a viral promoter. In a preferred embodiment, the promoter is the T7 RNA polymerase promoter. The regulatory components, e.g., the transcription promoter, can vary among nucleic acids at different addresses of the plurality. For example, different promoters can be used to vary the amount of polypeptide produced at different addresses.

In one embodiment, the nucleic acid also includes at least one site for recombination, e.g., homologous recombination or site-specific recombination, e.g., a lambda att site or variant thereof; a lox site; or a FLP site. In a preferred embodiment, the recombination site lacks stop codons in the reading frame of a nucleic acid encoding a test amino acid sequence. In another preferred embodiment, the recombination site includes a stop codon in the reading frame of a nucleic acid encoding a test amino acid sequence.

In another embodiment, the nucleic acid includes a sequence encoding a cleavage site, e.g., a protease site, e.g., a site cleaved by a site-specific protease (e.g., a thrombin site, an enterokinase site, a PreScission site, a factor Xa site, or a TEV site), or a chemical cleavage site (e.g., a methionine, preferably a unique methionine (cleavage by cyanogen bromide) or a proline (cleavage by formic acid)).

The nucleic acid can include a sequence encoding a second polypeptide tag in addition to the affinity tag. The second tag can be C-terminal to the test amino acid sequence and the affinity tag can be N-terminal to the test amino acid sequence; the second tag can be N-terminal to the test amino acid sequence, and the affinity tag can be C-terminal to the test amino acid sequence; the second tag and the affinity tag can be adjacent to one another, or separated by a linker sequence, both being N-terminal or C-terminal to the test amino acid sequence. In one embodiment, the second tag is an additional affinity tag, e.g., the same or different from the first tag. In another embodiment, the second tag is a recognition tag. For example, the recognition tag can report the presence and/or amount of test polypeptide at an address. Preferably the recognition tag has a sequence other than the sequence of the affinity tag. In still another embodiment, a plurality of polypeptide tags (e.g., less than 3, 4, 5, about 10, or

about 20 tags) are encoded in addition to the first affinity tag. Each polypeptide tag of the plurality can be the same as or different from the first affinity tag.

The nucleic acid sequence can further include an identifier sequence, e.g., a non-coding nucleic acid sequence, e.g., one that is synthetically inserted , and allows for uniquely identifying the nucleic acid sequence. The identifier sequence can be sufficient in length to uniquely identify each sequence in the plurality; e.g., it is about 5 to 500, 10 to 100, 10 to 50, or about 10 to 30 nucleotides in length. The identifier can be selected so that it is not complementary or identical to another identifier or any region of each nucleic acid sequence of the plurality on the array.

The test amino acid sequence can further include a protein splicing sequence or intein. The intein can be inserted in the middle of a test amino acid sequence. The intein can be a naturally-occurring intein or a mutated intein.

The nucleic acid sequences encoding the test amino acid sequences can be obtained from a collection of full-length expressed genes (e.g., a repository of clones), a cDNA library, or a genomic library. The test amino acid sequences can be genes expressed in a tissue, e.g., a normal or diseased tissue. The test polypeptides can be mutants or variants of a scaffold protein (e.g., an antibody, zinc-finger, polypeptide hormone etc.). In yet another embodiment, the test polypeptides are random amino acid sequences, patterned amino acids sequences, or designed amino acids sequences (e.g., sequence designed by manual, rational, or computer-aided approaches). The plurality of test amino acid sequences can include a plurality from a first source, and plurality from a second source. For example, the test amino acid sequences on half the addresses of an array are from a diseased tissue or a first species, whereas the sequences on the remaining half are from a normal tissue or a second species.

In a preferred embodiment, each address of the plurality further includes one or more second nucleic acids, e.g., a plurality of unique nucleic acids. Hence, the plurality in toto can encode a plurality of test sequences. For example, each address of the plurality can encode a pool of test polypeptide sequences, e.g., a subset of a library or clone bank. A second array can be provided in which each address of the plurality of the second array includes a single or subset of members of the pool present at an address of the first array. The first and the second array can be used consecutively.

In other preferred embodiments, each address of the plurality further includes a second nucleic acid encoding a second amino acid sequence.

In one preferred embodiment, each address of the plurality includes a first test amino acid sequence that is common to all addresses of the plurality, and a second test amino acid sequence that is unique among all the addresses of the plurality. For example, the second test amino acid sequences can be query sequences whereas the first amino test amino acid sequence can be a target sequence. In another preferred embodiment, each address of the plurality includes a first test amino acid sequence that is unique among all the addresses of the plurality, and a second test amino acid sequence that is common to all addresses of the plurality. For example, the first test amino acid sequences can be query sequences whereas the second amino test amino acid sequence can be a target sequence. The second nucleic acid encoding the second test amino acid sequence can include a sequence encoding a recognition tag and/or an affinity tag.

At at least one address of the plurality, the first and second amino acid sequences can be such that they interact with one another. In one preferred embodiment, they are capable of binding to each other. The second test amino acid sequence is optionally fused to a detectable amino acid sequence, e.g., an epitope tag, an enzyme, a fluorescent protein (e.g., GFP, BFP, variants thereof). The second test amino acid sequence can be itself detectable (e.g., an antibody is available which specifically recognizes it). The method can further include detecting the second test amino acid sequence at each address of the plurality, e.g., by detecting the detectable amino acid sequence (e.g., the epitope tag, enzyme or fluorescent protein).

In another preferred embodiment, one is capable of modifying the other (e.g., making or breaking a bond, preferably a covalent bond, of the other). For example, the first amino acid sequence is kinase capable of phosphorylating the second amino acid sequence; the first is a methylase capable of methylating the second; the first is a ubiquitin ligase capable of ubiquitinating the second; the first is a protease capable of cleaving the second; and so forth. The method can further include detecting the modification at each address of the plurality.

These embodiments can be used to identify an interaction or to identify a compound that modulates, e.g., inhibits or enhances, an interaction.

The binding agent can be attached to the substrate. For example, the substrate can be derivatized and the binding agent covalent attached thereto. The binding agent can be attached via a bridging moiety, e.g., a specific binding pair. (e.g., the substrate

contains a first member of a specific binding pair, and the binding agent is linked to the second member of the binding pair, the second member being attached to the substrate).

In yet another embodiment, an insoluble substrate (e.g., a bead or particle), is disposed at each address of the plurality, and the binding agent is attached to the insoluble substrate. The insoluble substrate can further contain information encoding its identity, e.g., a reference to the address on which it is disposed. The insoluble substrate can be tagged using a chemical tag, or an electronic tag (e.g., a transponder). The insoluble substrate can be disposed such that it can be removed for later analysis.

In another aspect, the invention features a method of providing an array across a network, e.g., a computer network, or a telecommunications network. The method includes: providing a substrate comprising a plurality of addresses, each address of the plurality having a binding agent; providing a plurality of nucleic acid sequences, each nucleic acid sequence comprising a sequence encoding a test amino acid sequence and an affinity tag that is recognized by the binding agent; providing on a server a list of either (i) nucleic acid sequences of the plurality or (ii) subsets of the plurality (e.g., sets of randomized sequences); transmitting the list across a network to a user; receiving at least one selection of the list from the user; disposing the one or more nucleic acid sequence corresponding to the selection on an address of the plurality; and providing the substrate to the user. In one embodiment, the plurality of nucleic acid sequences includes a random segment, e.g., a segment encoding a randomized polypeptide sequence.

In one embodiment, each nucleic acid sequence is disposed at a unique address. For example, if a subset is selected, each nucleic acid sequence of the subset is disposed at a unique address. In another embodiment, a plurality of nucleic acid sequences are disposed at each address.

The method can further include contacting each address of the plurality with one or more of (i) a transcription effector, and (ii) a translation effector. Optionally, the substrate is maintained under conditions permissive for the amino acid sequence to bind the binding agent. One or more addresses can then be washed, e.g., to remove at least one of (i) the nucleic acid, (ii) the transcription effector, (iii) the translation effector, and/or (iv) an unwanted polypeptide, e.g., an unbound polypeptide or unfolded polypeptide. The array can optionally be contacted with a compound, e.g., a chaperone; a protease; a protein-modifying enzyme; a small molecule, e.g., a small

organic compound (e.g., of molecular weight less than 5000, 3000, 1000, 700, 500, or 300 Daltons); nucleic acids; or other complex macromolecules e.g., complex sugars, lipids, or matrix molecules.

The array can be further processed, e.g., prepared for storage. It can be enclosed in a package, e.g., an air- or water-resistant package. The array can be desiccated, frozen, or contacted with a storage agent (e.g., a cryoprotectant, an anti-bacterial, an anti-fungal). For example, an array can be rapidly frozen after being optionally contacted with a cryoprotectant. This step can be done at any point in the process (e.g., before or after contacting the array with an RNA polymerase; before or after contacting the array with a translation effector; or before or after washing the array). The packaged product can be supplied to a user with or without additional contents, e.g., a transcription effector, a translation effector, a vector nucleic acid, an antibody, and so forth.

In a preferred embodiment, each test amino acid sequence in the plurality of addresses is unique. For example, a test amino acid sequence can differ from all other test amino acid sequence of the plurality by 1, or more amino acid differences, (e.g., about 2, 3, 4, 5, 8, 16, 32, 64 or more differences; and, by way of example, has about 800, 256, 128, 64, or 32, 16, 8, 4, or fewer differences). In another preferred embodiment, the test amino acid sequence encoded by the nucleic acid at each address of the plurality is identical to all other test amino acid sequences in the plurality of addresses. In a preferred embodiment, the affinity tag encoded by the nucleic acid at each address of the plurality is the same, or substantially identical to all other affinity tags in the plurality of addresses. In another preferred embodiment, the nucleic acid at each address of the plurality encodes more than one affinity tag. In yet another preferred embodiment, the affinity tag encoded by the nucleic acid at an address of the plurality differs from at least one other affinity tag in the plurality of addresses.

In a preferred embodiment, the affinity tag is fused directly to the test amino acid sequence, e.g., directly amino-terminal, or directly carboxy-terminal. In another preferred embodiment, the affinity tag is separated from the test amino acid by one or more linker amino acids, e.g., 1, 2, 3, 4, 5, 6, 8, 10, 12, 20, 30 or more amino acids, preferably about 1 to 20, or about 3 to 12 amino acids. The linker amino acids can include a cleavage site, flexible amino acids (e.g., glycine, alanine, or serine, preferably

glycine), and/or polar amino acids. The linker and affinity tag can be amino-terminal or carboxy-terminal to the test amino acid sequence.

The nucleic acid can be a RNA, or a DNA (e.g., a single-stranded DNA, or a double stranded DNA). In a preferred embodiment, the nucleic acid includes a plasmid DNA or a fragment thereof; an amplification product (e.g., a product generated by RCA, PCR, NASBA); or a synthetic DNA.

The nucleic acid can further include one or more of: a transcription promoter; a transcription regulatory sequence; a untranslated leader sequence; a sequence encoding a cleavage site; a recombination site; a 3' untranslated sequence; a transcriptional terminator; and an internal ribosome entry site. In one embodiment, the nucleic acid sequence includes a plurality of cistrons (also termed "open reading frames"), e.g., the sequence is dicistronic or polycistronic. In another embodiment, the nucleic acid also includes a sequence encoding a reporter protein, e.g., a protein whose abundance can be quantitated and can provide an indication of the quantity of test polypeptide fixed to the plate. The reporter protein can be attached to the test polypeptide, e.g., covalently attached, e.g., attached as a translational fusion. The reporter protein can be an enzyme, e.g., β-galactosidase, chloramphenicol acetyl transferase, β-glucuronidase, and so forth. The reporter protein can produce or modulate light, e.g., a fluorescent protein (e.g., green fluorescent protein, variants thereof, red fluorescent protein, variants thereof, and the like), and luciferase.

The transcription promoter can be a prokaryotic promoter, a eukaryotic promoter, or a viral promoter. In a preferred embodiment, the promoter is the T7 RNA polymerase promoter. The regulatory components, e.g., the transcription promoter, can vary among nucleic acids at different addresses of the plurality. For example, different promoters can be used to vary the amount of polypeptide produced at different addresses.

In one embodiment, the nucleic acid also includes at least one site for recombination, e.g., homologous recombination or site-specific recombination, e.g., a lambda att site or variant thereof; a lox site; or a FLP site. In a preferred embodiment, the recombination site lacks stop codons in the reading frame of a nucleic acid encoding a test amino acid sequence. In another preferred embodiment, the recombination site includes a stop codon in the reading frame of a nucleic acid encoding a test amino acid sequence.

In another embodiment, the nucleic acid includes a sequence encoding a cleavage site, e.g., a protease site, e.g., a site cleaved by a site-specific protease (e.g., a thrombin site, an enterokinase site, a PreScission site, a factor Xa site, or a TEV site), or a chemical cleavage site (e.g., a methionine, preferably a unique methionine (cleavage by cyanogen bromide) or a proline (cleavage by formic acid)).

The nucleic acid can include a sequence encoding a second polypeptide tag in addition to the affinity tag. The second tag can be C-terminal to the test amino acid sequence and the affinity tag can be N-terminal to the test amino acid sequence; the second tag can be N-terminal to the test amino acid sequence, and the affinity tag can be C-terminal to the test amino acid sequence; the second tag and the affinity tag can be adjacent to one another, or separated by a linker sequence, both being N-terminal or C-terminal to the test amino acid sequence. In one embodiment, the second tag is an additional affinity tag, e.g., the same or different from the first tag. In another embodiment, the second tag is a recognition tag. For example, the recognition tag can report the presence and/or amount of test polypeptide at an address. Preferably the recognition tag has a sequence other than the sequence of the affinity tag. In still another embodiment, a plurality of polypeptide tags (e.g., less than 3, 4, 5, about 10, or about 20 tags) are encoded in addition to the first affinity tag. Each polypeptide tag of the plurality can be the same as or different from the first affinity tag.

The nucleic acid sequence can further include an identifier sequence, e.g., a non-coding nucleic acid sequence, e.g., one that is synthetically inserted, and allows for uniquely identifying the nucleic acid sequence. The identifier sequence can be sufficient in length to uniquely identify each sequence in the plurality; e.g., it is about 5 to 500, 10 to 100, 10 to 50, or about 10 to 30 nucleotides in length. The identifier can be selected so that it is not complementary or identical to another identifier or any region of each nucleic acid sequence of the plurality on the array.

The test amino acid sequence can further include a protein splicing sequence or intein. The intein can be inserted in the middle of a test amino acid sequence. The intein can be a naturally-occurring intein or a mutated intein.

The nucleic acid sequences of the plurality can be obtained from a collection of full-length expressed genes (e.g., a repository of clones), a cDNA library, or a genomic library. The test amino acid sequences can be genes expressed in a tissue, e.g., a normal or diseased tissue. The test polypeptides can be mutants or variants of a

scaffold protein (e.g., an antibody, zinc-finger, polypeptide hormone etc.). In yet another embodiment, the test polypeptides are random amino acid sequences, patterned amino acids sequences, or designed amino acids sequences (e.g., sequence designed by manual, rational, or computer-aided approaches). The plurality of test amino acid sequences can include a plurality from a first source, and plurality from a second source. For example, the server can be provided with lists of test amino acid sequences associated with a diseased tissue or a first species in addition to lists of test amino acid sequences associated with a normal tissue or a second species.

The binding agent can be attached to the substrate. For example, the substrate can be derivatized and the binding agent covalent attached thereto. The binding agent can be attached via a bridging moiety, e.g., a specific binding pair. (e.g., the substrate contains a first member of a specific binding pair, and the binding agent is linked to the second member of the binding pair, the second member being attached to the substrate).

In yet another embodiment, an insoluble substrate (e.g., a bead or particle), is disposed at each address of the plurality, and the binding agent is attached to the insoluble substrate. The insoluble substrate can further contain information encoding its identity, e.g., a reference to the address on which it is disposed. The insoluble substrate can be tagged using a chemical tag, or an electronic tag (e.g., a transponder). The insoluble substrate can be disposed such that it can be removed for later analysis.

The invention also features a computer system including (i) a server storing a list of amino acid sequences and/or their descriptors, and (ii) software configured to: (1) send a list of amino acid sequence and/or their descriptors to a client; (2) receive from the client a plurality of selected amino acid sequences from the list ; and (3) interface with an array provider (e.g., a robotic system, or a technician) so as to dispose on a substrate nucleic acids encoding the selected amino acid sequences, each at a plurality of addresses.

The invention also features a computer system including (i) a server storing a list of amino acid sequences and/or their descriptors, and (ii) software configured to: (1) receive information (e.g., from a client, e.g. a remote client) about interactions between the amino acid sequences and a sample (e.g., a sample including an unknown); (2) compare the information about the interactions to a database of interactions observed for other samples (e.g., other unknowns or other controls), and (3) send results of the comparison to a user (e.g., the client).

The term "randomized" refers to one or more sequences in which any subunit (e.g., nucleotide, ribonucleotides, or amino acid) can be present at one, more than one or all specified or unspecified positions; therefore, for such positions as are randomized, the sequence of the finished molecule is not pre-determined, but is left to at least some degree of chance. A process of randomizing a protein or nucleic acid can refer to a synthetic method in which the incorporation of a subunit is left to at least some degree of chance.

In one embodiment, the user uses a plurality of protein ligands (e.g., random ligands on an array) to rapidly detect the presence of an agent that is wholly or partially composed of macromolecules (virus, bacteria, parasite, cancer antigen, disease markers, etc.) in a sample derived from a patient (e.g., from blood, urine, exhaled air, etc.)or in an environment (air, water supply, etc.). Exemplary situations for rapidly diagnosis include situations in which: (1)the patient is recently admitted to a hospital with a fever and other signs of infection, a bacterial or viral agent is suspected, but a specific diagnosis is required; (2)a group of individuals are traveling in a foreign country and become ill with a common set of symptoms suspected of being a possible infection, the identity of the agent is not known and the ability to rapidly detect the infection and identify the agent becomes essential; (3)an air or water supply is suspected of contamination by a bioterrorist agent, detection of this agent becomes essential. In another example, the specific and rapid diagnosis of a particular variation of cancer will allow a more appropriate specifically-tailored chemotherapy. This invention will address these situations and will enable the appropriate detection of signature patterns that point to a specific diagnosis (or detection)in a field environment and in real time.

Virtually all pathogens (and diseases)have several or more (usually many) macromolecules that are unique to the pathogen (or disease)that are not normally found in healthy blood (or other tested sources). In addition, infected (or affected)hosts produce response proteins that could also signal disease. Detection of these marker proteins enables identification of the pathogen (or specifically diagnose the disease). The use of marker proteins has already been demonstrated in a number of cases for example: HBSAg (a protein indicating the presence of active type B hepatitis), p24 (used to detect the presence of HIV), CA-125 (used to detect ovarian cancer though also found in some lung cancers), CMV Antigen (Cytomegalovirus

detection), Cryptococcal Antigen (detection of cryptococcal infection), Rheumatoid Factor (rheumatoid arthritis), etc.

Some advantages of some embodiments described herein include: 1) there is no need to select or identify proteins or macromolecules that are specific enough to point to a particular pathogen or disease, and 2) more than one protein associated with a particular pathogen or disease can be detected by a profiling method. In addition, some antigens are associated with more than one disease. For example, CA-125 is unusual in healthy women and is often elevated in ovarian cancer. But the same antigen is also elevated in some lung cancers. Thus, integration of additional information should improve the specificity of diagnosis.

In one embodiment, instead of identifying a single specific protein that is diagnostic of the pathogen (or disease), a signature pattern of several or more proteins is identified on a test array and this is used to make the diagnosis. One implementation uses a set of identical arrays comprising a collection of specific capture probes (e.g., each with a unique binding property). For example, the capture probe can have a chemistry or structure that has a high likelihood of binding to macromolecules. The identity of each element in the array is known, but its binding specificity does not need to be known. The test array is probed with a specimen and the macromolecules in the specimen will bind to various elements of the test array. The test array is then examined for a signature pattern that specifically differentiates between individuals infected with the agent and normal individuals. The pattern for the signature can be determined heuristically by training the test array on test sets and then testing unknowns. For example, fuzzy logic, genetic algorithms, or multi-dimensional distant metrics can be used to compare signatures or profiles, e.g., to classify profiles as related or unrelated and so forth.

Thus an advantage of this method is that the "marker proteins " need not ever be identified, as long as the specific elements to which they bind on the test array can be replicated. Moreover, because it is a signature pattern of several macromolecules binding to elements of the array that makes the diagnosis, it can significantly increase the sensitivity and specificity of the approach.

Some embodiments include the following:

1. Test array -This array contains many elements to which macromolecules will bind. Each element in the array can be identified well enough to

reproducibly place it on an array whenever desired. A collection of arrays with the same elements will be needed for the training set. The elements in the array can be varied to allow a broad range of binding specificities. The choice of test arrays can be adjusted depending on the application. This inventions works particularly well in conjunction with NAPPA, which allows the simple adjustment of a protein array to include any desired protein elements by simply spotting different samples of DNA. Examples of possible arrays: a. Pathogen proteome array. Among the best arrays in this context would be a collection of proteins present in the targeted pathogen. Because most macromolecules in an organism will bind to other proteins in the organism, there will be a high hit rate and many spots will light. All proteins in the proteome are not required, just a large sample. b. Host proteome. Another good choice will be a large collection of host proteins. Because the pathogen interacts with the host, a collection of host proteins that interact with pathogen macromolecule can be used.c. Collection of random proteins. There is enough variation in protein chemistry that a well-randomized collection of proteins or peptides will bind some fraction of the pathogen macromolecules and create a signature pattern. d. Small molecules -a well randomized set of small molecules with varied chemistries could also be used here

2. Detection system -This is a system that detects binding of any macromolecule to any element of the array. The detection system need not require the detection of any specific protein, it can merely detect that something has bound to certain elements. This can be accomplished in several ways, not limited to the following: a. Surface Plasmon Resonance -a change in index of refraction is detected, which indicates macromolecule binding. b. Surface Plasmon Enhanced Illumination -a resonance is set up by an array of holes or features. A change of index of refraction at the binding surface shifts resonant wavelength and demonstrates macromolecule binding. c. Sample labeling -The macromolecules in the sample are labeled with fluorescent or radioactive markers. The detector measures the presence of the marker at specific positions on the microarray and indicates that a macromolecule has bound.

3. Samples -Samples can be acquired from affected individuals and from control normal individuals. Enough samples are acquired to provide an opportunity train the algorithms to differentiate a normal sample from an affected sample.

An exemplary Process (see attached figure)

1. A collection of identical test arrays are created. 2. Samples from individuals with or without the pathogen are each reacted with a test array. 3. Test arrays are appropriately washed to eliminate non-specific binding. 4. Raw data are collected on each test array showing the elements with specific binding 5. Heuristic analysis compares infected to normal individuals to find specific patterns 6. Several specific patterns may be expected: a. Constant background -patterns illuminated in all samples b.. Protein Signature -patterns illuminated only in infected samples c. Individual variation -elements that light in some individuals and not others 7. Complete a statistical analysis to find those elements with good predictability in evaluating unknowns. 8. Unknown samples are then read and compared to the determined patterns to make diagnoses.

Some advantages of certain aspects may include: 1. The signature pattern does not need to be a single absolute pattern, only a set of patterns that statistically indicate the presence of the pathogen 2. The same set of arrays may be used to identify more than one pathogen 3. This tool can be used to find a pattern for a pathogen, even if the pathogen has never before been identified. A group of affected individuals can be identified and a separate group of control individuals can be identified, the heuristic algorithms can be used to find a specific pattern 4. In one embodiment, the method does not use of mass spectrometry to identify the proteins bound to the array. Detection of binding could be done with relatively simple instruments that could be fairly compact in size.

5. Some pathogens have the ability to mutate and change their phenotype. In some embodiments, the continued inclusion of new data into the training sets enables the signature patterns can evolve with the changing pathogens. Thus the arrays may never become obsolete. 6. Once clear patterns are identified by research, the complexity of the arrays can be reduced to just those elements that show a good positive predictive value and good negative predictive value -creating simple arrays that can be deployed for field use.

Expected future uses and/or commercial applications may include: 1. Clinical diagnosis of patients with suspected infections of known pathogens. 2. Clinical diagnosis of specific variants of pathogens for rapid adjustment of antibiotic therapy 3. Rapid clinical diagnosis of patients with other disorders where marker proteins may be helpful, e.g., specific forms of cancer, specific types of rheumatic diseases, acute

myocardial infarction 4. Clinical evaluation of populations with suspected pathogen of unknown character (requires a control population of unaffected individuals). 5. Evaluation of samples (water supply, food, air, etc.)for the presence of microorganisms or toxic macromolecules. 6. Evaluation of samples (e.g., water supply, food, air, etc.) for the presence of a threat, e.g., bioterrorist contamination.

All cited patents, patent applications, and references are incorporated herein in their entirety for all purposes. In particular, U.S. Published Application 2002-0192673 is incorporated by reference.

Brief Description of the Drawing

FIG. 1, 2, and 3 are an exemplary method for evaluating a sample. In FIG. 1 samples are contacted to replicate arrays. Samples from infected individuals produce profiles illustrated schematically in 1a, 1b, . . . and 1e. Corresponding samples from normal individuals are illustrated in 2a, 2b, ... and 2e. In FIG. 2, the profiles are compared, e.g., using heuristic analysis to identify locations common to all individuals (horizontal hatching), specific to infected individuals, or specific to normal individuals (vertical hatching) or without correlation (diagonal hatching). In. Fig. 3, the patterns are summarized.

Detailed Description of Embodiments

Macromolecular arrays of proteins can be used to detect an interaction profile (or binding signature pattern) for a biological sample specimen, e.g., to detect a pathological condition in a subject

The invention provides in various embodiments methods to enable the user to rapidly detect the presence of an agent by identifying macromolecules such as proteins from a virus, a bacterium, a fungus, a parasite, a cancer antigen, etc., in a biological sample specimen from a patient (for example, blood, urine, perspiration, amniotic fluid, lachrymal secretions, vaginal secretions, semen, exhaled air, saliva, sweat, cerebrospinal fluid, tears, feces, or extracts of cells or tissue) or in an environment (for example, air, water supply, soil, vegetation, etc.).

A variety of public health situations would benefit from an ability to more rapidly diagnose the presence of a specific pathogen, including: (1) a patient is recently

admitted to a hospital with fever and other signs of infection, a bacterial or viral agent is suspected, and a specific diagnosis is required; (2) a group of individuals is traveling in a foreign country and become ill with a common set of symptoms suspected of infectious etiology, the identity of the agent is not known, and the ability to rapidly detect and identify the agent of infection is essential; (3) a group of dead livestock or wild animals are found, and distinction between a pathogen or another causative factor must be made; (4) an air or water supply is suspected of contamination by a bioterrorist infectious agent, and rapid detection and identification of this agent become essential.

There also exist a number of situations where specific and rapid diagnosis of a cancer variant would allow a more rapid implementation of an appropriate specifically-tailored chemotherapy. Current diagnoses may use procedures such as analyzing samples following bronchoscopy, needle biopsy, endoscopy, surgical biopsy, CAT, MRI and PET scans, the procedures may be performed serially, so that more than a month can elapse merely to diagnose the nature and stage possible metastasis of the cancer, prior to the onset of a therapeutic intervention. Performance of procedures in this manner can be limited by availability of equipment, even in a major medical center. Aspects of this disclosure address these situations, providing methods of detection of appropriate signature patterns that point to a specific diagnosis or detection in a clinical setting or in a field environment, and enabling availability of diagnostic data, e.g., in a matter of hours.

In one embodiment, instead of identifying a single specific protein that is considered to be uniquely diagnostic of the pathogen (or disease), a profile (or signature pattern) is identified on a test array that includes a plurality of proteins, and this profile or signature is then used to make the diagnosis. In one embodiment, the test array is replicable, i.e., it is one of a set of identical arrays comprising a collection of specific binding sites, each site having a unique binding chemistry likely to bind to at least one macromolecule, such as a protein, peptide or oligopeptide. The identity of the source of each element in the array is known, as are reproducible methods of obtaining and applying each element, but neither the specific identity (function, sequence, etc.) nor its binding specificity need be known. Elements of known identity can be included.

A "sample" of each of the replicable test arrays is probed with biological specimens. The method thus uses a plurality of duplicated test arrays, or can re-use samples of arrays, or can use a combination. Components present in the specimen bind

to various individual proteins positioned at "addressable" locations (the locations being replicable for each "sample" of the test array) on the test array. Each test array is then examined for a signature pattern that differentiates between specimens, for example, from contrasting sets of individuals, for example, between a set of individuals infected with a pathogenic agent and a set of uninfected individuals. The signature pattern for each target such as a pathogen is determined heuristically, by training the test array on a set of known positive specimens, and then testing unknown specimens. See Figure 1.

An advantage of this method is that it is not necessary to identify and analyze the chemistry of any of the proteins in the array, or to use known proteins, including positive marker proteins, as long as the specific proteins in a signature pattern to which specimen molecules bind on the test array are replicable, that is, can be replicated to the same addressable location for each of the additional plurality of samples of the arrays. Each sample of the test array is then identical with respect to components present of addressable locations.

Moreover, because the diagnosis is comprised of a signature pattern of binding of several components of the specimen to locations of the array, the sensitivity of the approach is significantly increased compared to use of a single marker. Once a signature binding pattern has been identified, an array of less complexity, i.e., having fewer addressable locations, can be produced for each diagnostic application.

The following components are involved in various embodiments of the invention. "Test array" is in general a two-dimensional substrate which contains a plurality of binding components, such as proteins, each at an addressable location; a subset of the components of certain characteristic locations can bind molecules in a biological sample specimen.

A positionally addressable array can comprise a plurality of different substances, for example proteins, polypeptide, peptide or oligopeptide molecules comprising functional domains of the proteins, protein containing cellular material, or even whole cells or viruses, on a solid support (or substrate). A test array comprises from about 5 to about 1,000 locations, or about 50 to about 5,000 locations, or about 100 to about 10,000 locations. Each component in the array is identified only sufficiently well to reproducibly deposit it at the addressable location in a consistent quantity, and to affix it to the location on the array under scale-up conditions of production required for the large numbers of arrays for commercial use.

Proteins can be affixed to the substrate, for example, to an aldehyde treated glass slide (MacBeath *et al.*, Science 289: 1760, 2000). A plurality of arrays can be used as a "training set." The number and variety of protein components in the array should be sufficient to allow a broad range of specimen component binding specificities and can be subsequently reduced in second generation arrays for a specific diagnostic application. The choice of source and number of components to deposit in a test array can thus be adjusted for the application. In one embodiment, a NAPPA ("nucleic acid programmed protein arrays"), which enables a protein array to include a plurality of protein elements by spotting different samples of DNA is used. A "nucleic acid programmable protein array" or "NAPPA" refers to an array having a plurality of nucleic acids disposed at addressable locations on the array, on which synthesis of a polypeptide encoded by the nucleic acid is conducted such that the polypeptide remains bound to the array.

Examples of potential sources of compositions for arrays include, all or part of a pathogen proteome array, or a host proteome array. Also suitable are a random protein array, and a small molecule array. A pathogen proteome array or subset, *i.e.*, a collection of proteins present in a targeted pathogen, is a preferred embodiment because many macromolecules in a target organism will bind other proteins in the same organism with high affinity. Therefore, for detection of a signature pattern from a specimen of an infected subject, a high hit rate will be obtained, and many spots can be detectably bound. A complete set of proteins in the proteome is not required, as a large subset of pathogen proteins provides an initial array sufficient for the training set.

A host proteome, or subset thereof, i.e., a large collection of host proteins such as human proteins is affixed to the substrate, each protein at an addressable location, is another preferred embodiment. Because a pathogen of necessity interacts with a cognate host, a collection of host proteins on an array will enable identification of a signature pattern of bound proteins that interact with pathogen proteins. Methods of obtaining and rapidly purifying a set of proteins and preparing glass slides with an array of these proteins are known (Zhu et al., 2001 Science 293: 2101-2105); protein-protein interactions resulting from binding to proteins on such arrays can be analyzed, for example, with fluorescent dyes (MacBeath et al. 2000 Science 289: 1760-1763). Hosts of pathogens are not limited to animals, and can also be crop plants such as wheat, corn, soy bean, oat and crop vegetables and fruit.

Collections of random coding nucleic acids or proteins can be prepared based on saturation mutagenesis (U.S. patent number 6,171,820) or other known techniques (U.S. patent number 6,361,974; U.S. patent application 2002-0048772). These random proteins provide enough variation that proteins or peptides will bind some components of a pathogen specimen and establish thereby a signature pattern. Small molecules – a well randomized set of small molecules with varied chemistries, such as are available from ChemBridge, Corp. (San Diego, CA), PharmaCore (High Point), or the NIH, could also be used. It is also possible to select out non-identical members from a random collection or to select out a representative set of members from a random collection for efficient array production. Conversely, it is possible to use a random collection without sampling it.

The term "detectably bound" means that the presence of a component from the target specimen can be detected bound to an addressable location on the array, and indicates that a component of the specimen has bound. A "detection system" is a system that detects binding of a macromolecule to a protein at any location on the array. The detection method need not require detection of the presence of a specific protein or interaction, rather the method detects that a composition has bound to an addressable location in an array. Detection can be accomplished in several ways, including Surface Plasmon Resonance, in which a change in index of refraction is detected as a result of macromolecule binding. A change of index of refraction at the binding surface can be detected by Surface Plasmon Enhanced Illumination, in which a resonance is set up by an array of holes or features due to shifts in resonant wavelength which demonstrate macromolecule binding. Sample labeling can be used, *i.e.*, macromolecules in the sample are labeled with one or more fluorescent or radioactive markers.

Specimen samples can include, e.g., a biological fluid, cell or tissue, or an environmental sample, in the case of positive controls are acquired from affected subjects, for example, subjects having an infection or a cancer, in contrast to control unaffected individuals, or from an environment. A plurality of samples and control specimens can be used to provide for statistically significant training of the algorithms to differentiate a normal sample from an affected sample. The plurality can be, for example, at least 2, or 3 to 10 samples, or 5 to 12, or 50-200 samples for any particular target, disease or disorder.

Test arrays allow the direct analysis of discrete protein binding and other activities without the complications of adverse in vivo effects. For example, a low-density (96 well format) protein array has been developed in which proteins, spotted onto a nitrocellulose membrane and biomolecular interactions, were visualized by autoradiography (Ge, H. 2000 *Nucleic Acids Res.* 28:e3, I-VII). In another example, a high-density protein array (100,000 samples within 222 X 222 mm) that was used for antibody screening was formed by spotting proteins onto polyvinylidene difluoride (PVDF;Lueking *et al.* 1999 *Anal. Biochem.* 270:103-111). Proteins have been printed on a flat glass plate that contained wells formed by an enclosing hydrophobic Teflon mask, and the arrayed antigens were detected using enzyme-linked immunosorbent assay (ELISA) techniques (Mendoza *et al.* (1999) *Biotechniques* 27:778-788.). A large-scale *in vitro* analysis of biochemical activity using affinity-purified yeast proteins has been performed in the context of an array of 6144 yeast strains, each bearing a plasmid expressing a different GST-ORF fusion (Martzen *et al.* 1999 Science *286*, 1153-1155). Proteins have been covalently linked to chemically derivatized flat glass slides in a high-density array (1600 spots per square centimeter), and protein-protein and protein-small molecule interactions were detected by fluorescence or radioactive decay (MacBeath and Schreiber (2000) *Science* 289:1760-1763). A high-density array of 18,342 bacterial clones has been generated, each expressing a different single-chain antibody, for screening antibody-antigen interactions (De Wildt *et al.* (2000) *Nature Biotech.* 18:989-994).

The binding component is in one embodiment attached to the substrate of the test array. For example, the substrate can be derivatized and the binding component covalent attached thereto. The binding agent can be attached via a bridging moiety, e.g., a specific binding pair. (e.g., the substrate contains a first member of a specific binding moiety, and the binding component is linked to the second member of the binding pair, the second member being attached to the substrate). Alternatively, the term "test array," as used herein can also refer to a set of micro-wells with a plurality of addresses in which the binding components are deposited.

A database, e.g., as a computer memory or a computer readable medium of the collection of signatures for each test array, can be included. The database can have a field representing a result (e.g., a qualitative or quantitative result). The database includes a record for each address of the plurality present on the array. The records can

be clustered or have a reference to other records (e.g., including hierarchical groupings) based on the result.

In one embodiment, each test location in the plurality of addresses features a composition that is unique. Alternatively, a portion of the addresses can be redundant, providing internal controls. Redundancy is controlled by knowledge of the complexity of the components, e.g., a proteome, a random protein library, etc. For example, a proteome having about 35,000 unique gene products, if represented by a test array of 5,000 products, has little redundancy.

In another embodiment, a test array having components from a target organism or a disease cell at addressable locations can be used to identify an interaction or to identify a compound that modulates, e.g., inhibits or enhances, an interaction.

In another embodiment, a polypeptide or protein at an addressable location includes a cleavage site, e.g., a protease site, e.g., a site cleaved by a site-specific protease (e.g., a thrombin site, an enterokinase site, a PreScission site, a factor Xa site, or a TEV site), or a chemical cleavage site (e.g., a methionine, preferably a unique methionine (cleavage by cyanogen bromide) or a proline (cleavage by formic acid)). The test amino acid sequence can further include a protein splicing sequence or intein. The intein can be inserted in the middle of a test amino acid sequence. The intein can be a naturally-occurring intein or a mutated intein.

A variety of test amino acid sequences can be disposed at different addresses of the plurality. For example, the test array can include proteins that are expressed in a tissue, e.g., a normal or diseased tissue. In yet another embodiment, the test polypeptides are random amino acid sequences, patterned amino acids sequences, or designed amino acids sequences (e.g., sequence designed by manual, rational, or computer-aided approaches). The proteins can include a plurality from a first source, and plurality from a second source. For example, the proteins on half of the addresses of an array are from a diseased tissue or a first species, whereas the sequences on the remaining half are from a normal tissue or a second species.

An address of the test array can further include one or a plurality of additional polypeptides. For example, an address can include a pool of test polypeptides, e.g., a subset of polypeptides encoded by a library or clone bank. A second test array can be provided in which an address of the plurality of the second test array includes a single

or subset of members of the pool present at an address of the first array. The first and the second test arrays can be used simultaneously or consecutively.

In one embodiment, each address of the plurality includes a first test amino acid sequence that is common to addresses of the plurality, and a second test component that is unique among the addresses of the plurality. For example, the second test component can be query compositions whereas the first amino test amino acid sequence can be a target sequence.

A test array can be stored for use at a later time, for example, can be rapidly frozen after being optionally contacted with a cryoprotectant. The packaged product can be supplied to a user with or without additional components.

Proteins at addressable locations can be obtained from a collection of full-length expressed genes (e.g., a repository of clones), for example, expressed in a tissue, e.g., a normal or diseased tissue.

The method can further include washing the substrate, e.g., after sufficient contact with a specimen. The wash step can be repeated, e.g., one or more times, e.g., until an excess of a component is removed. The wash step can remove unbound proteins. The stringency of the wash step can vary, e.g., the salt, pH, and buffer composition of the wash buffer can vary. For example, the substrate can be washed with a chaotrope, (e.g., guanidinium hydrochloride, or urea). In a subsequent step, the chaotrope can itself be washed from the array, and the compositions can be renatured. Alternatively, contacting the specimen can be performed under conditions of sufficient stringency that only limited washing is necessary prior to continuing with the method.

The method can further include contacting the substrate with a second substrate. For example, in an embodiment wherein the substrate is a gel, the gel can be contacted with a second gel, and the contents of one gel can be transferred to another (e.g., by diffusion or electrophoresis).

The addressable locations can have a composition further containing an epitope (e.g., recognized by a monoclonal antibody), or a binding agent (e.g., avidin or streptavidin, GST, or chitin binding protein). Detection can entail contacting each address of the plurality with a binding agent, e.g., a labeled biotin moiety, labeled glutathione, labeled chitin, a labeled antibody, etc. In another embodiment, each address of the plurality is contacted with an antibody specific to an amino acid sequence. The antibody can be labeled, e.g., with a fluorophore.

Kits provided herein can further include a database, e.g., in computer memory or a computer readable medium (e.g., a CD-ROM, a magnetic disc, flash memory). Each record of the database can include a descriptor or reference for the physical location of the signature pattern on the array. The records can be clustered or have a reference to other records (e.g., including hierarchical groupings) based on the result.

The kit can also include instructions for use of the test array, or a link or indication of a network resource (e.g., a web site) having instructions for use of the arrays or the above database of records describing the addresses of the signature patterns for each application.

In another aspect, the invention provides a method of providing an array across a network, e.g., a computer network, or a telecommunications network. The method includes transmitting across a network to a user; receiving at least one selection of the list from the user; transmitting at least one signature patterns corresponding to the selection of an application; and providing the substrate to the user.

The invention can include a computer system including a server storing a list of test array or their descriptors, and software configured to: send a list of test arrays and/or their descriptors to a client; receive from the client one or a plurality of applications desired for synthesis, or selected from the list; and interface with an array provider (e.g., a robotic system, or a technician) so as to dispose on a substrate proteins or other compositions, each at a plurality of addresses.

The term “address,” as referred to herein, is a positionally distinct portion of a substrate in an array. Thus, a component at a first address can be positionally distinguished from a component at a second address. The address is located in and/or on the substrate or in micro wells. The address can be distinguished by two coordinates (e.g., x-y) in embodiments using two-dimensional arrays, or by three coordinates (e.g., x-y-z) in embodiments using three-dimensional arrays or multiple.

The term “substrate,” as used herein in the context of arrays (as opposed to a substrate of an enzyme), refers to a composition in or on which a set of protein polypeptides, or small molecules is disposed. The substrate may be discontinuous. An illustrative case of a discontinuous substrate is a set of gel pads separated by a partition.

As used herein, the terms “peptide,” “polypeptide,” and “protein” are used interchangeably. Generally, these terms refer to polymers of amino acids which are at least three amino acids in length.

"Unique reagent" refers to a component that differs from other components at other addresses within the plurality of addresses. (An array can include additional pluralities of addresses in addition to the plurality being described; a plurality can include, e.g., at least 10, 100, or 1000 addresses). The component can differ from the components at other addresses in terms of recognition and binding of one or more different specimens. A unique component can be a molecule, e.g., a biological macromolecule (e.g., a protein, a polypeptide, or a carbohydrate), or a small organic compound. In the case of biological polymers, a structural difference can be a difference in sequence at least one position. In addition, a structural difference, e.g., for polymers having the same sequence, can be a difference in conformation (e.g., due to allosteric modification; meta-stable folding; alternative native folded states; prion or prion-like properties) or a modification (e.g., covalent and non-covalent modifications (e.g., a bound ligand))

Substrates.

Both solid and porous substrates are suitable for recipients for the encoding nucleic acids described herein. A substrate material can be selected and/or optimized to be compatible with the spot size (e.g., density) required for the application.

In one embodiment, the substrate is a solid substrate. Potentially useful solid substrates include: mass spectroscopy plates (e.g., for MALDI), glass (e.g., functionalized glass, a glass slide, porous silicate glass, a single crystal silicon, quartz, UV-transparent quartz glass), plastics and polymers (e.g., polystyrene, polypropylene, polyvinylidene difluoride, poly-tetrafluoroethylene, polycarbonate, PDMS, acrylic), metal coated substrates (e.g., gold), silicon substrates, latex, membranes (e.g., nitrocellulose, nylon), and a glass slide suitable for surface plasmon resonance (SPR).

In another embodiment, the substrate is porous, e.g., a gel or matrix. Potentially useful porous substrates include: agarose gels, acrylamide gels, sintered glass, dextran, meshed polymers (e.g., macroporous crosslinked dextran, sephacryl, and sepharose), and so forth.

Substrates can have properties such as being opaque, translucent, or transparent. The addresses can be distributed, on the substrate in one dimension, e.g., a linear array; in two dimensions, e.g., a planar array; or in three dimensions, e.g., a three dimensional array. The solid substrate may be of any convenient shape or form, e.g., square,

rectangular, ovoid, or circular. In another embodiment, the solid substrate can be disc shaped and attached to a means of rotation.

In one embodiment, the substrate contains at least 1, 10, 100, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, or 10^9 or more addresses per cm^2. The center to center distance of each address can be at least about 5 mm, 1 mm, 100 μm, or can be less than about 10 μm, 1 μm, or 100 nm. The longest diameter of each address can be at least about 5 mm, 1 mm, or less than about 100 μm, 10 μm, 1 μm, or 100 nm. In one embodiment, each address contains at least about 1 μg, for example, 10 μg , or each address contains less than about 100 ng, 10 ng, 1 ng, 100 pg, 10 pg, 1 pg, or 0.1 pg of the protein. In another embodiment, each address contains at least about 100, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, or 10^9 or more molecules of the composition.

The substrate can be modified to facilitate the stable attachment of linkers, capture probes, or binding agents. Generally, a skilled artisan can use routine methods to modify a substrate in accordance with the desired application. The following are non-limiting examples of substrate modifications.

A surface can be amidated, e.g., by silylating the substrate, e.g., with trialkoxyaminosilane. Silane-treated surface can also be derivatized with homobifunctional and heterobifunctional linkers. The substrate can be derivatized, e.g., so it has a hydroxy, an amino (e.g., alkylamine), carboxyl group, N-hydroxy-succinimidyl ester, photoactivatable group, sulfhydryl, ketone, or other functional group available for reaction. The substrates can be derivatized with a mask in order to only derivatized limited areas; a chemical etch or UV light can be used to remove derivatization from selected regions.

For preparation of glass slides, it is possible to derivatize the individual spots, or to derivatize the entire slide then use a physical mask, chemical etch, or UV light to cover or remove the derivatization in the areas between spots.

Substrates can be partitioned. In one preferred embodiment, each address is partitioned from all other addresses in order to maintain separation of molecules at each of the addresses. The substrate can be partitioned, e.g., by depressions, grooves, or photoresist. For example, the substrate can be a microchip with microchannels and reservoirs etched therein, e.g., by photolithography. Other non-limiting examples of substrates include multi-welled plates, e.g., 96-, 384-, 1536-, 6144- well plates, and polyclimethyl siloxane (PDMS) plates. Such high-density plates are commercially

available, often with specific surface treatments. Depending on the optimal volume required for each application, an appropriate density plate is selected. In another embodiment, the partitions are generated by a hydrophobic substance, e.g., a Teflon mask, grease, or a marking pen (e.g., from Snowman, Japan).

In one embodiment, the substrate is designed with reservoirs isolated by protected regions, e.g., a layer of photoresist. A mask can be focused or placed on the substrate, and a photoresist barrier separating the two reservoirs can be removed by illumination. The method can also include moving the substrate in order to facilitate mixing.

Substrates can have a planar surface. In another embodiment, the addresses are not physically partitioned, but diffusion is limited on the planar substrate, e.g., by increasing the viscosity of the solution, by providing a matrix with small pore size which excludes large macromolecules, and/or by tethering at least one of the aforementioned macromolecules. In yet another embodiment, modest or even substantial diffusion to neighboring addresses is permitted. Results, e.g., a signal of a label, are processed, e.g., using a computer system, in order to determine the position of the center of the signal. Thus, by compensating for radial diffusion, the address can be accurately determined.

Substrates are not limited to two-dimensional. A three-dimensional substrate can be generated, e.g., by successively applying layers of a gel matrix on a substrate. Each layer contains a plurality of addresses. The porosity of the layers can vary, e.g., so that alternating layers have reduced porosity.

In another embodiment, a three-dimensional substrate includes stacked two-dimensional substrates, e.g., in a tower format. Each two-dimensional substrate is accessible to a dispenser and detector.

A substrate can be a micromachined chip. Chips are made with glass and plastic materials, using rectangular or circular geometry. Wells and fluid channels are machined into the chip, and then the surfaces are derivatized. A humidity-controlled chamber can be used to control evaporation.

A disk geometry (also termed "CD format") is another suitable substrate for the microarray. Sample addition and reactions are performed while the disk is spinning (see PCT WO 00/40750; WO 97/21090; GB patent application 9809943.5; "The next small thing" (Dec. 9, 2000) *Economist Technology Quarterly* p. 8; PCT WO 91/16966; Duffy

et al. (1999) *Analytical chemistry;* 71, 20, (1999), 4669-4678). Thus, centrifugal force drives the flow of sample deposition of specimen and wash solutions.

The disc can include sample-loading areas, reagent-loading areas, reaction chambers, and detection chambers. Such microfluidic structures are arranged radially on the disc with the originating chambers located towards the disc center. Samples from a microtiter plate can be loaded using a liquid train and a piezo dispenser. Multiple samples can be separated in the liquid train by air gaps or an inert solution. The piezo dispenser then dispenses each sample onto appropriate application areas on the CD surface, e.g., a rotating CD surface. The volume dispensed can vary, e.g., less than about 10 pL, 50 pL, 100 pL, 500 pL, 1 nL, 5 nL, or 50 nL. After entry on the CD, the centripetal force conveys the dispensed sample into appropriate reaction chambers. Flow between chambers can be guided by barriers, transport channels, and/or surface interactions (e.g., between the walls and the solution). The depth of channels and chambers can be adjusted to control volume and flow rate in each area.

A master CD can be made by deep reactive ion etching (DRIE) on a 6-inch silicon wafer. This master disk can be plated and used as a model to manufacture additional CDs by injection molding (e.g., ´ mic AB, Uppsala, Sweden). A stroboscope can be used to synchronize the detector with the rotation of the CD in order to track individual detection chambers.

Components of the test array or of the specimen sample can have an affinity tag An amino acid sequence that encodes a member of a specific binding pair can be used as an affinity tag. The other member of the specific binding pair is attached to the substrate, either directly or indirectly.

One class of specific binding pair is a peptide epitope and the monoclonal antibody specific for it. Any epitope to which a specific antibody is or can be made available can serve as an affinity tag. See Kolodziej and Young (1991) *Methods Enz.* 194:508-519 for general methods of providing an epitope tag. Exemplary epitope tags include HA (influenza haemagglutinin; Wilson *et al.* (1984) *Cell* 37:767), myc (e.g., Mycl-9E10, Evan *et al. (1985) Mol. Cell. Biol.* 5:3610-3616), VSV-G, FLAG, and 6-histidine (see, e.g., German Patent No. DE 19507 166).

An antibody can be coupled to a substrate of an array, e.g., indirectly using *Staphylococcus aureus* protein A, or streptococcal protein G. The antibody can be covalently bound to a derivatized substrate, e.g., using a crosslinker, e.g., N-hydroxy-

succinimidyl ester. The test polypeptides with epitopes such as Flag, HA, or myc are bound to antibody-coated plates.

Another class of specific binding pair is a small organic molecule or simple polymer, and a polypeptide sequence that specifically binds it. Specific binding pairs include glutathione and glutathione-S-transferase, chitin binding protein and chitin, cellulase and cellulose, methotrexate and dihydrofolate reductase, and FK506 and FKBP.

Art-known methods of tethering components such as proteins, e.g., the use of specific binding pairs, are suitable for the affinity or chemical capture of polypeptides on the array. Appropriate substrates include commercially available streptavidin and avidin-coated plates, for example, 96-well Pierce Reacti-Bind Metal Chelate Plates or Reacti-Bind Glutathione Coated Plates (Pierce, Rockford, IL). Histidine- or GST-tagged test polypeptides are immobilized on either 96-well Pierce Reacti-Bind Metal Chelate Plates or Reacti-Bind Glutathione Coated Plates, respectively, and unbound proteins are optionally washed away. Yet another class of specific binding pair is a metal, and a polypeptide sequence which can chelate the metal. An exemplary pair is Ni^{2+} and the hexa-histidine sequence (see U.S. Patent No. 4,877,830; 5,047,513; 5,284,933; and 5,130,663.).

An affinity tag can be a dimerization sequence, e.g., a homodimerization or heterodimerization sequence., preferably a heterodimerization sequence. In one illustrative example, the affinity tag is a coiled-coil sequence, e.g., the heptad repeat region of Fos. The binding agent coupled to the array is the heptad repeat region of Jun. The test polypeptide is tethered to the substrate by heterodimization of the Fos and Jun heptad repeat regions to form a coiled-coil.

In another embodiment, the affinity tag is provided by an unnatural amino acid, e.g., with a side chain having functional properties different from a naturally occurring amino acid. The binding agent attached to the substrate functions as a chemical handle to either bind or react with the affinity tag.

In a related embodiment, the affinity tag is a free cysteine which can be oxidized with a thiol group attached to the substrate to create a disulfide bond that tethers the test polypeptide.

Recognition Tags

A variety of recognition tags can be used. For example, an epitope to which an antibody is available can be used as a recognition tag. The tag can be located at the N- or C-terminal to the sequence of interest. The tag is recognized, e.g., directly, or indirectly (e.g., by binding of an antibody).

For Green fluorescent protein, coding regions of interest are taken from the FLEX repository and transferred into fusion vectors encoding either an N- or C-terminal green fluorescent protein (GFP) tag. Complexes are detected by fluorescence spectroscopy (Spectra Max Gemini, Molecular Devices). The environment of a fluorophore has a strong effect on the quantum yield of fluorescence (i.e., the ratio of emitted to absorbed photons) through collisional processes and resonance energy transfer (a radiative process), and the concentration of target-query complexes that gives an acceptable signal-to-noise ratio is determined experimentally. Conventional fluorescence spectroscopy and fluorescence polarization methods can be used to detect protein-protein interactions. See, e.g., Garcia-Parajo *et al.* (2000) Proc. Natl. Acad. Sci. USA *97*, 7237-7242.

For enzymatic reporters, horseradish peroxidase (HRP) or alkaline phosphatase (AP) polypeptide sequences can be used as a recognition tag. The addition of chromogenic substrate and subsequent colorimetric readout allows for the ready detection of the retention of the second polypeptide. Luciferase can be used as a recognition tag as described in U.S. Patent No. 5,641,641.

MS (Mass Spectroscopy) recognition is achieved by analysis by mass spectroscopy, e.g., MALDI-TOF, which is indicative of the presence of a bound polypeptide.

A patient specimen is contacted to a sample of the test the array. Non-limiting examples of patient samples include serum proteins, proteins extracted from a biopsy obtained from the patient, and so forth as described herein. In addition, cells or cell extracts can be contacted to the array in order to query for components displayed on the cell surface.

In one embodiment, the specimen is modified with a compound prior to being contacted to the array. For example, the components in the specimen can be biotinylated. Addresses that bind proteins in the specimen are then identified by contacting the array with labeled streptavidin or labeled avidin. In another

embodiment, the sample is unlabeled. MALDI, SPR, or another techniques are used to identify if a protein is bound at each address. Arrays can be designed to identify proteins associated with various pathologies, e.g., to detect antigens associated with cancer at various stages (for example, early, pre-metastatic stages or late stage cancer) or to provide a prediction (for example, to quantitate the abundance of an antigen correlated with a condition). The subject can be a human patient, an animal, a forensic sample, or an environmental sample (e.g., from a waste system).

Detection of binding of a test sample macromolecule to one or more addressable locations can be achieved also by labeling the macromolecules in the test sample with a label which is radioactive, or a fluorophore, or a chemical, an epitope (to be identified by a specific antibody), or by labeling with a nucleic acid (to be amplified and identified for example by a labeled complementary nucleic acid for hybridization). Such labeling of the test sample macromolecules can be achieved chemically, for example, via an -SH group of a cysteine residue.

Transcription Effectors

RNA-directed RNA polymerases and DNA-directed RNA polymerases are both suitable transcription effectors.

DNA-directed RNA polymerases include bacteriophage T7 polymerase , phage T3, phage φII, Salmonella phage SP6, or Pseudomonas phage gh-l, as well as archeal RNA polymerases, bacterial RNA polymerase complexes, and eukaryotic RNA polymerase complexes.

T7 polymerase is a preferred polymerase. It recognizes a specific sequence, the T7 promoter (see e.g., U.S. Patent No. 4,952,496), which can be appropriately positioned upstream of an encoding nucleic acid sequence. Although, a DNA duplex is required for recruitment and initiation of T7 polymerase, the remainder of the template can be single stranded. In embodiments utilizing other RNA polymerases, appropriate promoters and initiations sites are selected according to the specificity of the polymerase.

RNA-directed RNA polymerases can include Qβ replicase, and RNA-dependent RNA polymerase.

Translation Effectors

In one embodiment, the transcription/translation mix is in a minimal volume, and this volume is optimized for each application. The volume of translation effector at each address can be less than about 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, or 10^{-9} L. During dispensing and incubation, the array can be maintained in an environment to prevent evaporation, e.g., by covering the wells or by maintaining a humid atmosphere.

In another embodiment, the entire substrate can be coated or immersed in the translation effector. One possible translation effector is a translation extract prepared from cells. The translation extract can be prepared e.g., from a variety of cells, e.g., yeast, bacteria, mammalian cells (e.g., rabbit reticulocytes), plant cells (e.g., wheat germ), and archebacteria. In a preferred embodiment, the translation extract is a wheat germ agglutinin extract or a rabbit reticulocyte lysate. In another preferred embodiment, the translation extract also includes a transcription system, e.g., a eukaryotic, prokaryotic, or viral RNA polymerase, e.g., T7 RNA polymerase. In a preferred embodiment, the translation extract is disposed on the substrate such that it can be removed by simple washing. The translation extract can be supplemented, e.g., with additional amino acids, tRNAs, tRNA synthases, and energy regenerating systems. In one embodiment, the translation extract also include an amber, ochre, or opal suppressing tRNA. The tRNA can be modified to contain an unnatural amino acid. In another embodiment, the translation extract further includes a chaperone, e.g., an agent which unfolds or folds polypeptides, (e.g., a recombinant purified chaperones, e.g., heat shock factors, GroEL/ES and related chaperones, and so forth. In another embodiment, the translation extract includes additives (e.g., glycerol, polymers, etc.) to alter the viscosity of the extract.

Affinity Tags

An amino acid sequence that encodes a member of a specific binding pair can be used as an affinity tag. The other member of the specific binding pair is attached to the substrate, either directly or indirectly.

One class of specific binding pair is a peptide epitope and the monoclonal antibody specific for it. Any epitope to which a specific antibody is or can be made available can serve as an affinity tag. See Kolodziej and Young (1991) *Methods Enz.* 194:508-519 for general methods of providing an epitope tag. Exemplary epitope tags include HA (influenza haemagglutinin; Wilson *et al.* (1984) *Cell* 37:767), myc (e.g.,

Myc1-9E10, Evan *et al.* (1985) *Mol. Cell. Biol.* 5:3610-3616), VSV-G, FLAG, and 6-histidine (see, e.g., German Patent No. DE 19507 166).

An antibody can be coupled to a substrate of an array, e.g., indirectly using *Staphylococcus aureus* protein A, or streptococcal protein G. The antibody can be covalently bound to a derivatized substrate, e.g., using a crosslinker, e.g., N-hydroxy-succinimidyl ester. The test polypeptides with epitopes such as Flag, HA, or myc are bound to antibody-coated plates.

Another class of specific binding pair is a small organic molecule, and a polypeptide sequence that specifically binds it. See, for example, the specific binding pairs listed in Table 1.

Table 1

Protein	Ligand
glutathione-S-transferase,	glutathione
chitin binding protein	chitin
Cellulase (CBD)	cellulose
maltose binding protein	amylose, or maltose
dihydrofolate reductases	methotrexate
FKBP	FK506

Additional art-known methods of tethering proteins, e.g., the use of specific binding pairs are suitable for the affinity or chemical capture of polypeptides on the array. Appropriate substrates include commercially available streptavidin and avidin-coated plates, for example, 96-well Pierce Reacti-Bind Metal Chelate Plates or Reacti-Bind Glutathione Coated Plates (Pierce, Rockford, IL). Histidine- or GST-tagged test polypeptides are immobilized on either 96-well Pierce Reacti-Bind Metal Chelate Plates or Reacti-Bind Glutathione Coated Plates, respectively, and unbound proteins are optionally washed away.

In one embodiment, the polypeptide is an enzyme, e.g., an inactive enzyme, and ligand is its substrate. Optionally, the enzyme is modified so as to form a covalent bond with its substrate. In another embodiment, the polypeptide is an enzyme, and the ligand is an enzyme inhibitor.

Yet another class of specific binding pair is a metal, and a polypeptide sequence which can chelate the metal. An exemplary pair is Ni^{2+} and the hexa-histidine sequence (see U.S. Patent No. 4,877,830; 5,047,513; 5,284,933; and 5,130,663.).

In still another embodiment, the affinity tag is a dimerization sequence, e.g., a homodimerization or heterodimerization sequence., preferably a heterodimerization sequence. In one illustrative example, the affinity tag is a coiled-coil sequence, e.g., the heptad repeat region of Fos. The binding agent coupled to the array is the heptad repeat region of Jun. The test polypeptide is tethered to the substrate by heterodimization of the Fos and Jun heptad repeat regions to form a coiled-coil.

In another embodiment (see also unnatural amino acids), the affinity tag is provided by an unnatural amino acid, e.g., with a side chain having functional properties different from a naturally occurring amino acid. The binding agent attached to the substrate functions as a chemical handle to either bind or react with the affinity tag.

In a related embodiment, the affinity tag is a free cysteine which can be oxidized with a thiol group attached to the substrate to create a disulfide bond that tethers the test polypeptide.

Disposal of Nucleic Acid Sequences on Arrays

The substrate and the liquid-handling equipment are selected with consideration for required liquid volume, positional accuracy, evaporation, and cross-contamination. The density of spots can depend on the liquid volume required for a particular application, and on the substrate, e.g., how much a liquid drop spreads on the substrate due to surface tension, and the positional accuracy of the dispensing equipment.

Numerous methods are available for dispensing small volumes of liquid onto substrates. For example, U.S. Patent No. 6,112,605 describes a device for dispensing small volumes of liquid. U.S. Patent No. 6,110,426 describes a capillary action-based method of dispensing known volumes of a sample onto an array.

Nucleic acid spotted onto slides can be allowed to dry by evaporation. Dry air can be used to accelerate the process.

Capture Probes. The substrate can include an attached nucleic acid capture probe at each address. In one aspect, capture probes can be used create a self-assembling array. A unique capture probe at each address selectively hybridizes to a

nucleic acid encoding a test amino acid sequence, thereby organizing each encoding nucleic acid to a unique address. The capture nucleic acid can be covalently attached or bound, e.g., to a polycationic surface on the substrate.

The capture probe can itself be synthesized in situ, e.g., by a light-directed method (see, e.g., U.S. Patent No. 5,445,934), or by being spotted or disposed at the addresses. The capture probe can hybridize to the nucleic acid encoding the test polypeptide. In a preferred embodiment, the capture probe anneals to the T7 promoter region of a single stranded nucleic acid encoding the test amino acid sequence. In another embodiment, the capture probe is ligated to the encoding nucleic acid sequence. In yet another embodiment, the capture probe is a padlock probe. In still another embodiment, the capture probe hybridizes to a nucleic acid encoding a test amino acid sequence, e.g., a unique region of the nucleic acid, or to a nucleic acid sequence tag provided on the nucleic acid for the purposes of identification.

Disposed Insoluble Substrates

One or more insoluble substrates having a binding agent attached can be disposed at each address of the array. The insoluble substrates can further include a unique identifier, such as a chemical, nucleic acid, or electronic tag. Chemical tags, e.g., such as those used for recursive identification in "split and pool" combinatorial syntheses. Kerr *et al.* (1993) *J. Am. Chem. Soc.*, 115:2529-2531) Nikolaiev *et al.* ((1993) *Peptide Res.* 6, 161-170) and Ohlmeyer *et al.*((1993) *Proc. Natl. Acad. Sci. USA* 90:10922-10926) describe methods for coding and decoding such tags. A nucleic acid tag can be a short oligonucleotide sequence that is unique for a given address. The nucleic acid tag can be coupled to the particle. In another embodiment, the encoding nucleic acid provides a unique identifier. The encoding nucleic acid can be coupled or attached to the particle. Electronic tags include transponders as mentioned below. The insoluble substrate can be a particle (e.g., a nanoparticle, or a transponder), or a bead.

Beads. The disposed particle can be a bead, e.g., constructed from latex, polystyrene, agarose, a dextran (sepharose, sephacryl), and so forth.

Transponders. U.S. Patent No. 5,736,332 describes methods of using small particles containing a transponder on which a handle or binding agent can be affixed. The identity of the particle is discerned by a read-write scanner device which can encode and decode data, e.g., an electronic identifier, on the particle (see also Nicolaou

et al. (1995) *Angew. Chem. Int. Ed. Engl.* 34:2289-2291*).* Test polypeptides are bound to the transponder by attaching to the handle or binding agent.

Disposed Nucleic acid Sequences

Any appropriate nucleic acid for translation can be disposed at an address of the array. The nucleic acid can be an RNA, single stranded DNA, a double stranded DNA, or combinations thereof. For example, a single-stranded DNA can include a hairpin loop at its 5' end which anneals to the T7 promoter sequence to form a duplex in that region. The nucleic acid can be an amplification products, e.g., from PCR (U.S. Patent No. 4,683,196 and 4,683,202); rolling circle amplification ("RCA," U.S. Patent No. 5,714,320), isothermal RNA amplification or NASBA (U.S. Patent Nos. 5,130,238; 5,409,818; and 5,554,517), and strand displacement amplification (U.S. Patent No. 5,455,166).

In one embodiment, the sequence of the encoding nucleic acid is known prior to being disposed at an address. In another embodiment, the sequence of the encoding nucleic acid is unknown prior to disposal at an address. For example, the nucleic acid can be randomly obtained from a library. The nucleic acid can be sequenced after the address on which it is placed has been identified as encoding a polypeptide of interest.

Amplification in situ

A nucleic acid disposed on the array can be amplified directly on the array, by a variety of methods, e.g., PCR (U.S. Patent No. 4,683,196 and 4,683,202); rolling circle amplification ("RCA," U.S. Patent No. 5,714,320), isothermal RNA amplification or NASBA , and strand displacement amplification (U.S. Patent No. 5,455,166).

Isothermal RNA amplification or "NASBA" is well described in the art (see, e.g., U.S. Patent Nos. 5,130,238; 5,409,818; and 5,554,517; Romano *et al.* (1997) *Immunol Invest.* 26:15-28; in technical literature for "RnampliFire™" Qiagen, CA). Isothermal RNA amplification is particularly suitable as reactions are homogenous, can be performed at ambient temperatures, and produce RNA templates suitable for translation.

Vectors for Expression

Coding regions of interest can be taken from a source plasmid, e.g., containing a full length gene and convenient restriction sites, or sites for homologous or site-

specific recombination, and transferred to an expression vector. The expression vector includes a promoter and an operably linked coding region, e.g., encoding an affinity tag, such as one described herein. The tag can be N or C terminal. The vector can carry a cap-independent translation enhancer (CITE, or IRES, internal ribosome entry site) for increased in vitro translation of RNA prepared from cloned DNA sequences. The fusion proteins will be generated with commercially available in vitro transcription/translation kits such as the Promega TNT Coupled Reticulocyte Lysate Systems or TNT Coupled Wheat Germ Extract Systems. Cell-free extracts containing translation component derived from microorganisms, such as a yeast, or a bacteria, can also be used.

In addition, the vector can include a number of regulatory sequences such as a transcription promoter; a transcription regulatory sequence; a untranslated leader sequence; a sequence encoding a protease site; a recombination site; a 3' untranslated sequence; a transcriptional terminator; and an internal ribosome entry site.

The vector or encoding nucleic acid can also include a sequence encoding an intein. Methods of using inteins for the regulated removal of an intervening sequence are described, e.g., in U.S. Patent Nos. 5,496,714 and 5,834,247. Inteins can be used to cyclize, ligate, and/or polymerize polypeptides, e.g., as described in Evans *et al.* (1999) *J Biol Chem* 274:3923 and Evans *et al.* (1999) *J Biol Chem* 274:18359.

Exemplary Useful Sequences

Useful sets of proteins for creating test arrays include naturally proteomic sets, randomized versions thereof, and artificial proteins (e.g., artificial variants of polypeptides that include a folded domain). Such proteins can be stored in a repository, see below. Proteins

Naturally occurring sequences. Naturally occurring sequences can be procured from cells of species from the kingdoms of animals, bacteria, archebacteria, plants, and fungi. Non-limiting examples of eukaryotic species include: mammals such as human, mouse (*Mus musculus*), and rat; insects such as *Drosophila melanogaster;* nematodes such as *Caernorhabditis elegans;* other vertebrates such as *Brachydanio rerio;* parasites such as *Plasmodium falciparum*, *Leishmania major*; fungi such as yeasts, *Histoplasma, Cryptococcus, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Pichia pastoris* and the like); and plants such as *Arabidoposis thaliana,* rice,

maize, wheat, tobacco, tomato, potato, and flax. Non-limiting examples of bacterial species include *E. coli, B. subtilis, Mycobacterium tuberculosis, Pseudomonas aeriginosa, Vibrio cholerae, Thermatoga maritime, Mycoplasma pneumoniae, Mycoplasma genitalium, Helicobacter pylori, Neisseria meningitidis,* and *Borrelia burgdorferi.* In additional, amino acid sequence encoded by viral genomes can be used, e.g., a sequence from rotavirus, hepatitis A virus, hepatitis B virus, hepatitis C virus, herpes virus, papilloma virus, or a retrovirus (e.g., HIV-1, HIV-2, HTLV, SIV, and STLV).

In a preferred embodiment, a cDNA library is prepared from a desired tissue of a desired species in a vector described herein. Colonies from the library are picked, e.g., using a robotic colony picker. DNA is prepared from each colony and used to program a NAPPA array.

Artificial sequences. The encoding nucleic acid sequence can encode artificial amino acid sequences. Artificial sequences can be randomized amino acid sequences, patterned amino acid sequence, computer-designed amino acid sequences, and combinations of the above with each other or with naturally occurring sequences. Cho *et al.* (2000) *J Mol Biol* 297:309-19 describes methods for preparing libraries of randomized and patterned amino acid sequences. Similar techniques using randomized oligonucleotides can be used to construct libraries of random sequences. Individual sequences in the library (or pools thereof) can be used to program a NAPPA array.

Dahiyat and Mayo (1997) *Science* 278:82-7 describe an artificial sequence designed by a computer system using the dead-end elimination theorem. Similar systems can be used to design amino acid sequences, e.g., based on a desired structure, such that they fold stably. In addition, computer systems can be used to modify naturally occurring sequences in order

Mutagenesis. The array can be used to display the products of a mutagenesis or selection. Examples of mutagenesis procedures include cassette mutagenesis (see e.g., Reidhaar-Olson and Sauer (1988) *Science* 241:53-7), PCR mutagenesis (e.g., using manganese to decrease polymerase fidelity), in vivo mutagenesis (e.g., by transfer of the nucleic acid in a repair deficient host cell), and DNA shuffling (see U.S. Patent No. 5,605,793; 5,830,721; and 6,132,970). Examples of selection procedures include complementation screens, and phage display screens. Mutagenic methods can be used to introduce randomization.

In addition, more methodical variation can be achieved. For example, an amino acid position or positions of a naturally occurring protein can be systematically varied, such that each possible substitution is present at a unique position on the array. For example, the all the residues of a binding interface can be varied to all possible other combinations. Alternatively, the range of variation can be restricted to reasonable or limited amino acid sets.

Collections. Additional collections include arrays having at different addresses one of the following combinations: combinatorial variants of a bioactive peptide; specific variants of a single polypeptide species (splice variants, isolated domains, domain deletions, point mutants); polypeptide orthologs from different species; polypeptide components of a cellular pathway (e.g., a signalling pathway, a regulatory pathway, or a metabolic pathway); and the entire polypeptide complement of an organism.

Informatics

A computer system, containing a repository of observed interaction is also featured. The computer system can be networked to receive data, e.g., raw data or processed data, from a data acquisition apparatus, e.g., a microchip slide scanner, a fluorescence microscope, or surface plasmon resonance.

The computer system includes a relational database. The database houses all data from multiple interaction profiles, e.g., using the same or different arrays. One table contains table rows for each array contacting evaluation, e.g., describing one or more the array production number, experiment date, array contents experimental conditions, and so forth. The raw data from an interaction microarray experiment, for example, is stored in a second table with table rows for each address on the array. This data includes the signature/profile information. The second table can have fields for observed fluorescence, background fluorescence, the amino acid sequences present at the microarray address, other annotations, links, cross-references and so forth.

Thus, the database provides a comprehensive catalog of profiles and information about specimens, samples, and/or controls. The system is designed to facilitate digital access to the data in order to interface the experimental results with predictive models of interactions. The system can be accessed in real time, e.g., as profile data is acquired, and from multiple network stations, e.g., multiple users within

a company (e.g., using an Intranet), multiple customers of a data provider (e.g., using secure Internet communication protocols), or multiple individuals across the globe (e.g., using the Internet).

Clustering algorithm can be applied to profiles in the database that are associated with particular information (e.g., a diagnosis, detection, species, etc.) Exemplary clustering algorithms include Eisen *et al.* ((1998) *Proc. Nat. Acad. USA* 95:14863) and Golub *et al.* ((1999) *Science* 286:531) for methods of clustering signatures/profiles. Other methods of comparing profiles include training a recognition algorithm. For example, the recognition algorithm can use one or more of: statistical analysis; fuzzy logic, hidden Markov models, regression, decision trees, neural networks, and genetic algorithms. See, variously, US published applications 2002-0146724; 2003-0059792; 2003-0049701; and 2003-0023385. Once trained, the information set can be used to send information that assigns a descriptor to incoming information to a user, e.g., a remote client.

EXAMPLES

The process in one embodiment includes: providing a plurality of identical test arrays, each array having a predetermined number of samples of proteins from a pathogen or an affected subject, the samples being identically arrayed at addressable locations; reacting each of the plurality of specimen biological samples from affected subjects and from unaffected controls with a test array; washing test arrays appropriately to eliminate non-specific binding; collecting raw data on each test array showing the pattern of elements that have specific binding of at least one component with a specimen sample; and analyzing heuristically to compare binding patterns of specimens of affected individuals to that of normal unaffected individuals, to find disease specific patterns. It is envisioned that many of the steps in these processes will be eliminated with further development. For example, once a disease-specific pattern has been identified, further analyzing is no longer required. Washing steps may be reduced or eliminated by establishing conditions that are sufficiently stringent that specific binding is obtained *ab initio*.

Several types of patterns are expected for binding of a specimen macromolecule to any given spot at an addressable location in an array. At some locations, a "constant

background" is observed, because the same patterns are illuminated in all specimen samples regardless of origin from an affected subject or a control. See Figure 1 for spots present in all individuals. At another set of locations, a "protein signature" of disease-specific "detectably bound" addressable locations, or illuminated patterns, is observed in samples from affected individuals only. See Figure 1. Finally, individual variation is observed as components that are illuminated in some individuals and not others, with no disease state correlation.

A statistical analysis of markably bound illuminated components at their addressable locations in with affected and unaffected control samples is performed, to find those components that correlate with a disease state with good predictability, for evaluating unknown samples. Unknown samples can then be read, and compared to the previously determined patterns, to provide diagnoses.

The signature pattern need not be a single absolute pattern. A set of patterns that statistically indicate the presence of the pathogen is sufficient. For example, each field in the pattern can have associated with it a variance or standard deviation. The same set of test arrays having components such as proteins at addressable locations may be used to identify more than one pathogen, if each pathogen has a reproducible pattern distinct from the pattern of other pathogens. An array can further be used to establish a pattern for a novel pathogen, even if the pathogen was not heretofore identified. As long as a first group of affected individuals and a second group of control individuals, or a first set of environmental samples having a target and a second set free of the target, can each be identified, the heuristic algorithms can be used to find a specific pattern. Embodiments of the invention do not require use of mass spectrometry to identify the proteins bound to the array. Detection of binding is accomplished with relatively simple instruments that are compact in size.

A number of pathogens, including HIV, influenza virus, and many protozoans, have the ability to mutate and change their expressed proteonic phenotype. Methods provided herein allow the continued inclusion of new data into the training sets, so that the signature patterns can evolve with the changing pathogens, and can maintain utility.

Once patterns are identified by the methods herein, the number of locations of the test array can be reduced to those that show a strong positive predictive value and good negative predictive value. This reduction creates simple or test arrays for an application, which can be more readily deployed for field use.

The methods herein are envisioned to provide commercial applications that include: clinical diagnosis of patients with suspected infections of known pathogens; clinical diagnosis of specific variants of pathogens for rapid adjustment of antibiotic therapy; rapid clinical diagnosis of patients with other disorders where marker proteins may be helpful, specific forms of cancer, specific types of autoimmune diseases such as rheumatic diseases, diseases such as acute myocardial infarction; clinical evaluation of human, animal or plant populations with suspected pathogen of unknown character (requires a control population of unaffected individuals); evaluation of samples (for example, water supply, food, air) for the presence of microorganisms or toxic macromolecules; and evaluation of samples for the presence of bioterrorist contamination. Food can be examined for the presence of a food poisoning agent such as a bacterium, and the bacterium can be identified by the methods herein, for example, the bacterium can be identified as a species of the genus *Salmonella* or the genus *Staphylococcus*.

What is claimed is:

1. A method for evaluating a specimen, the method comprising:

contacting a first sample to an array, wherein the first sample comprises a positive control for a target, the array comprising a plurality of capture probes affixed to a substrate at replicable locations wherein identification of compositions of the capture probes need not be known, thereby obtaining a first interaction profile,

contacting a second sample to the array or replicate thereof, the second sample not containing the target thereby obtaining a second interaction profile;

contacting a third sample of the array or replicate thereof, wherein the third sample is associated with a specimen or derived from a specimen to obtain a third interaction profile; and

comparing the third interaction profile with the first interaction profile, wherein presence of the target in the specimen is indicated by the presence of one or more features of the first interaction profile in the third interaction profile.

2. The method of claim 1, wherein each capture probe comprises a protein, and the proteins affixed to the substrate are from a mammal.

3. The method of claim 2, wherein the proteins affixed to the substrate are from a human.

4. The method of claim 2, wherein the proteins affixed to the substrate are from a cancer cell.

5. The method of claim 1, wherein each capture probe comprises a protein, and the proteins affixed to the substrate are from a pathogen.

6. The method of claim 5, wherein the pathogen is selected from the group of a virus, a bacterium, a fungus, and a protozoan.

7. The method of claim 1, wherein the target is a prion.

8. The method of claim 4, wherein the cell is selected from the group primary or metastatic cancers of lung, skin, leukemia, lymphoma, brain, breast, prostate, bowel, esophagus, liver, pancreas, and head and neck cancers.

9. The method of claim 6, wherein the bacterium is a genus selected from the group of *Actinobacillus*, *Bacillus*, *Borrelia*, *Brucella*, *Chlamydia*, *Clostridium*, *Coxiella*, *Enterococcus*, *Escherichia*, *Francisella*, *Hemophilus*, *Legionella*, *Mycobacterium*, *Neisseria*, *Pasteurella*, *Pseudomonas*, *Salmonella*, *Shigella*, *Staphylococcus*, *Streptococcus*, *Treponema*, and *Yersinia*.

10. The method of claim 9, wherein the *Bacillus* is *B. anthracis*.

11. The method of claim 9, wherein the *Mycobacterium* is *M. tuberculosis*.

12. The method of claim 9, wherein the *Borrelia* is *B. burgdorferi*.

13. The method of claim 1, wherein the proteins are from a spore of *Bacillus anthracis*.

14. The method of claim 6, wherein the virus is selected from the group of influenza, human immunodeficiency, Venezuelan equine encephalitis, West Nile, Lassa fever, hemorrhagic conjunctivitis, smallpox, rhinovirus, Lassa fever, Ebola, Rift Valley fever, Marburg, measles, mumps, yellow fever, herpes, hantavirus, hepatitis A, hepatitis B, hepatitis C, rotavirus, parvovirus, rabies, respiratory syncytial, rubella, Epstein Barr, Newcastle disease, hoof and mouth, tobacco mosaic, Glycine mosaic comovirus, and wheat American striate.

15. The method of claim 6, wherein the fungus is selected from the group of genera consisting of *Aspergillus*, *Candida*, *Phytophthora*, *Puccinia*, *Lichen*, and *Trichophyton*.

16. The method of claim 15, wherein the *Aspergillus* is *A. flavus*.

17. The method of claim 1, wherein the target comprises a bacterial or fungal toxin.

18. The method of claim 6, wherein the protozoan is *Plasmodium*, *Leishmania*, *Entamoeba*, *Enterocytozoan*, *Cryptosporidium*, and *Giardia*.

19. The method of claim 1, wherein the proteins affixed to the substrate are random proteins.

20. The method of claim 1, wherein the specimen is a biological fluid sample.

21. The method of claim 20, wherein the fluid is selected from the group of urine, saliva, lacrymal secretions, nasal discharge, blood, serum, plasma, lymph, perspiration, amniotic fluid, cerebrospinal fluid, ascites fluid, semen, vaginal secretions, feces, and cell extract.

22. The method of claim 1, wherein the specimen is an environmental sample.

23. The method of claim 22, wherein the environmental sample is selected from the group of soil suspension, air infusion, pond water, lake water, river water, ocean water, sewage, industrial effluent, food, beverages, consumable goods, packaged goods, mail, baggage, and fluid extract of a rubbing.

24. The method of claim 1, wherein method is re-iterated to obtain a statistically significant number of interaction profiles.

25. The method of claim 1, further comprising re-iterating the method to obtain an interaction profile for an additional target biological material.

26. The method of claim 25, wherein the signature binding pattern for the additional biological material is obtained using a fourth sample of the replicable test array.

27. The method of claim 1, wherein the first sample is associated with an unknown origin.

28. The method of claim 20, wherein the biological fluid is obtained from a patient with an acute medical condition.

29. The method of claim 29, wherein the acute medical condition is a cardiac condition.

30. The method of claim 29, wherein the cardiac condition is myocardial infarction or stroke.

31. The method of claim 20, wherein the biological fluid is obtained from a patient with an autoimmune disease.

32. The method of claim 31, wherein the autoimmune disease is selected from the group consisting of: multiple sclerosis, myasthenia gravis, Hashimoto's disease, systemic lupus erythematosis, uveitis, Guillain-Barre' syndrome, Grave's disease, idiopathic myxedema, autoimmune oophoritis, chronic immune thrombocytopenic purpura, colitis, diabetes, psoriasis, pemphigus vulgaris, and rheumatoid arthritis.

33. The method of claim 20, wherein the biological fluid is obtained from a patient with an inflammatory condition.

34. The method of claim 33, wherein the inflammatory condition is selected from the group consisting of asthma, allergy, and inflammatory bowel syndrome.

35. The method of claim 1, wherein in contacting the second sample of the array with the negative control, the signature for the target biological material additionally comprises at least one location present in the control pattern and absent from the target pattern.

36. A method comprising:

providing a plurality of nucleic acids, each nucleic acid of the plurality comprising a coding region comprising a random segment, the nucleic acids of the plurality being distributed on a plurality of substrates at addresseable locations;

producing polypeptides for each nucleic acid of the plurality by translating the coding region of each nucleic acid on each of the substrates;

contacting a plurality of samples to one of the substrates of the plurality;

producing a plurality of data profiles indicate interaction with each of the samples of the plurality with the polypeptides on each of the substrates.

37. The method of claim 36 wherein each of the profiles is associated with information describing the respective sample.

38. The method of claim 36 further comprising training a recognition algorithm to discriminate between one or more of the profile.

39. The method of claim 36 wherein the training comprises fuzzy logic, model building, or a genetic algorithm.

40. The method of claim 36 or 38 wherein the producing comprises in vitro translation on the substrate.

41. The method of claim 40 wherein the producing further comprises transcription on the substrate.

42. The method of claim 38 further comprising receiving information about a profile associated with an unknown, and determining an association for the profile.

43. The method of claim 42 further comprising sending information about the association to a user.

44. The method of claim 36 wherein providing a plurality of nucleic acids comprising synthesizing degenerate oligonucleotides.

45. A method comprising:

providing a database that comprises (1) a plurality of interaction profiles, wherein each interaction profile of the plurality describes interactions between a sample and a plurality of capture probes, and a characteristic of the sample and (2) an discriminatory function that enables an input interaction profile to be associated with one or more of the characteristics, wherein the plurality of capture probes comprises at least ten proteins that include a random segment.

46. A method comprising:

providing a database that comprises (1) a plurality of interaction profiles, wherein each interaction profile of the plurality describes interactions between a sample and a plurality of capture probes, and a characteristic of the sample and (2) an discriminatory function that enables an input interaction profile to be associated with one or more of the characteristics;

receiving, at a sever, information about an interaction profile associated with an unknown sample;

accessing the database and the discriminatory function to return characteristic information as a predictor of features of the unknown sample.

47. The method of claim 46 wherein the unknown sample comprises a patient sample, the patient being identified, but having an unknown disorder.

48. The method of claim 46 wherein the receiving comprises contacting the sample to a protein array that comprises the plurality of capture probes and receiving electronic information about the interaction of the sample and the protein array.

49. The method of claim 46 wherein the same plurality of capture probes is used for an unknown sample from a first and second patient, wherein the patients are of different ages, have different symptoms, or have different genetic backgrounds.

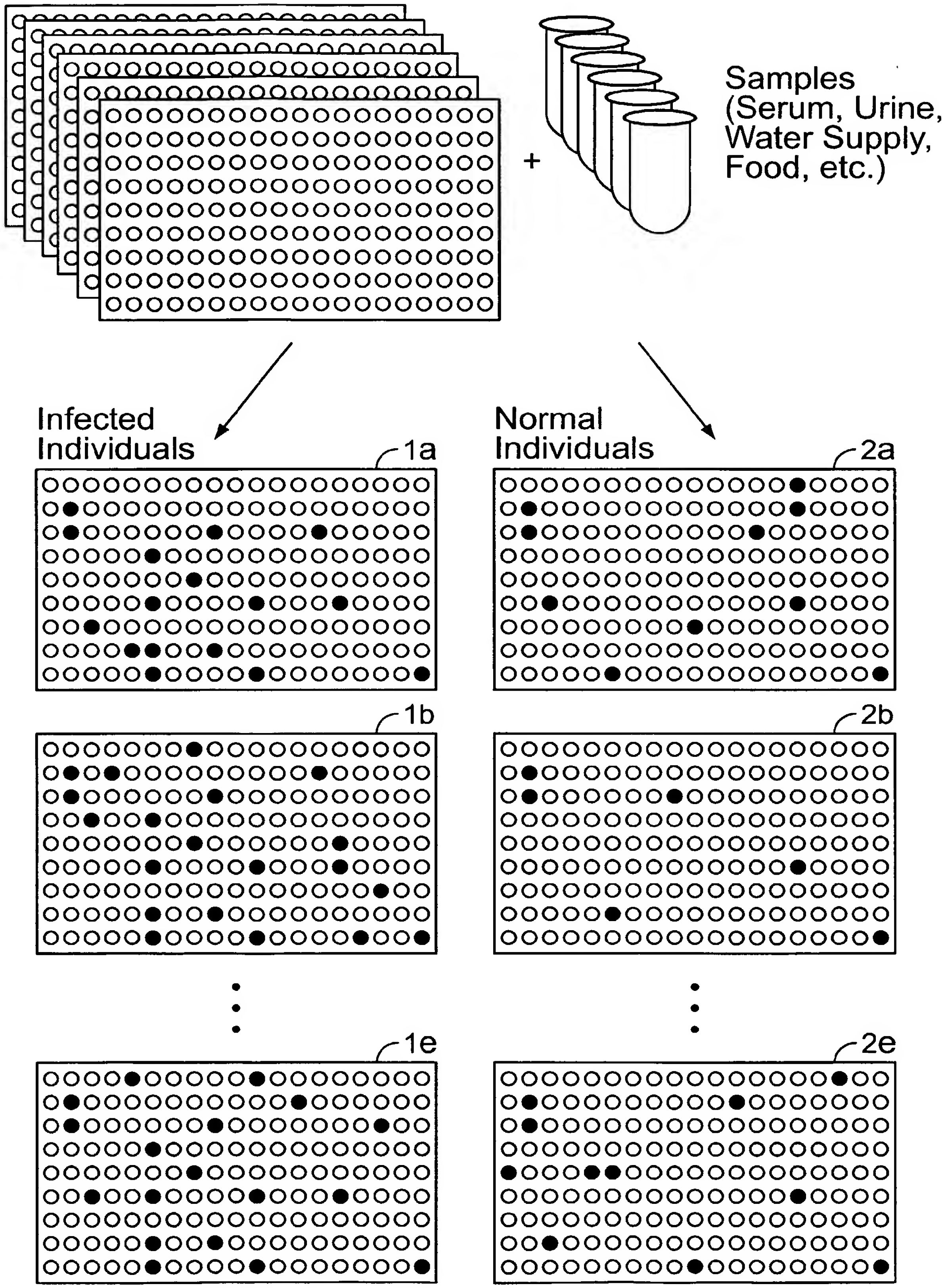

+
Samples (Serum, Urine, Water Supply, Food, etc.)
Infected Individuals
Normal Individuals
1a
2a
1b
2b
1e
2e

FIG. 1

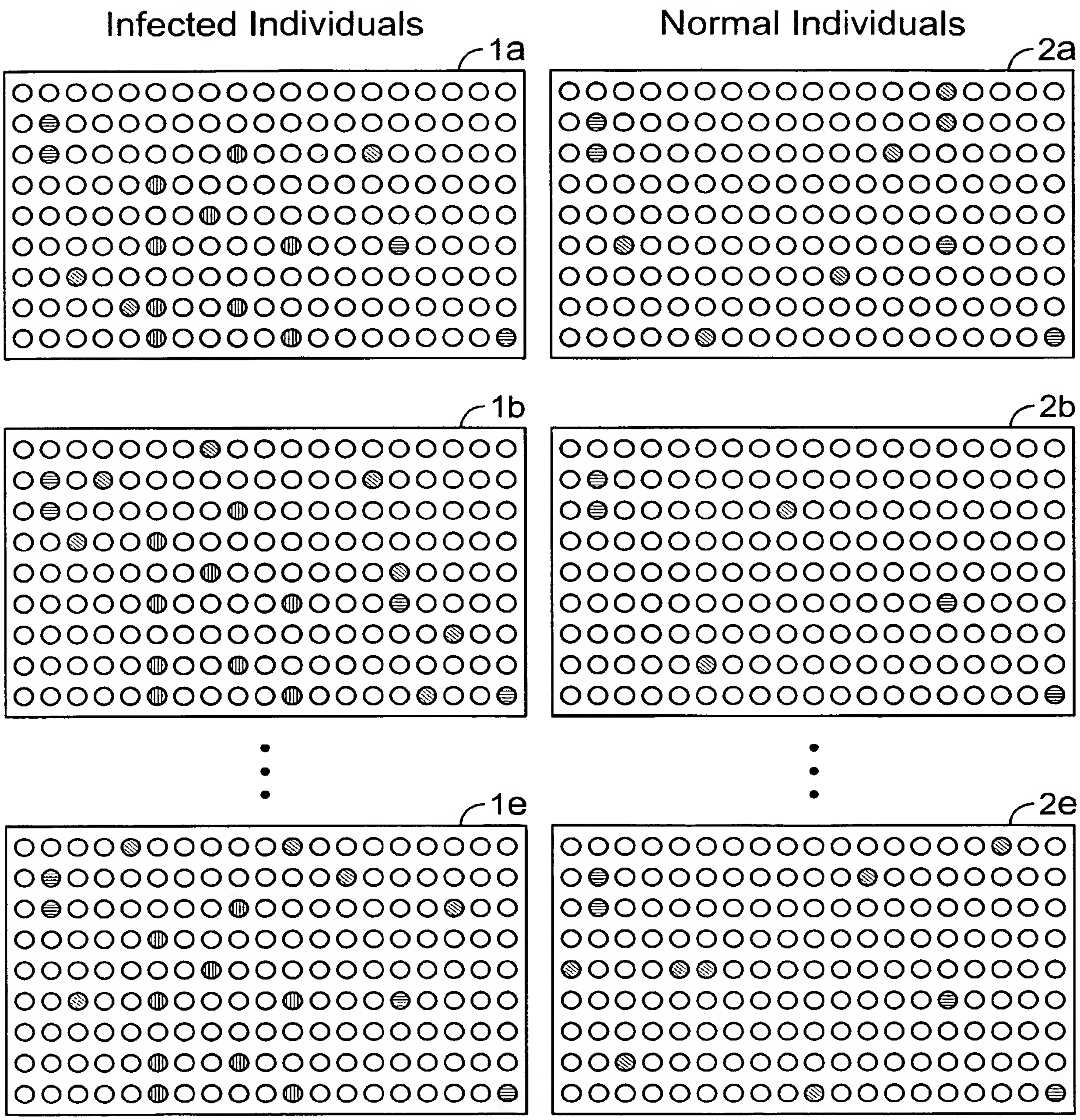


FIG. 2

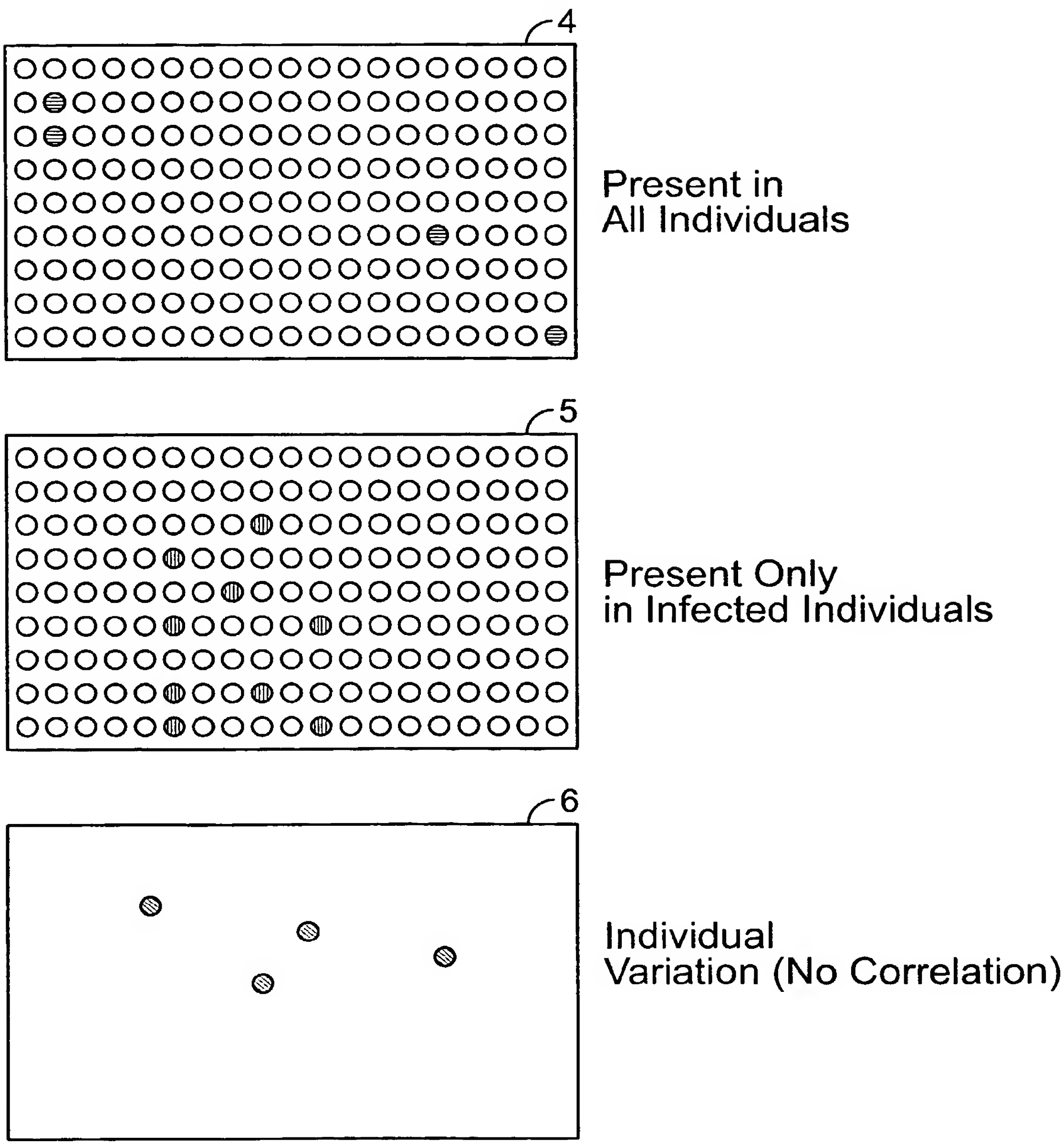

4
Present in
All Individuals
5
Present Only
in Infected Individuals
6
Individual
Variation (No Correlation)

FIG. 3

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
24 February 2005 (24.02.2005)

PCT

(10) International Publication Number
WO 2005/016230 A3

(51) International Patent Classification[7]: C12Q 1/68, C07H 21/02, 21/04

(21) International Application Number: PCT/US2003/017979

(22) International Filing Date: 9 June 2003 (09.06.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/387,034 7 June 2002 (07.06.2002) US

(71) Applicant (*for all designated States except US*): **PRESIDENT AND FELLOWS OF HARVARD COLLEGE** [US/US]; 17 Quincy Street, Cambridge, MA 02138-3876 (US).

(72) Inventor; and
(75) Inventor/Applicant (*for US only*): **LABAER, Joshua** [US/US]; 22 Moraine Street, Jamaica Plain, MA 02130-4316 (US).

(74) Agent: **MYERS, Louis**; Fish & Richardson P.C., 225 Franklin Street, Boston, MA 02110-2804 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

(88) Date of publication of the international search report: 26 January 2006

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 2005/016230 A3

(54) **Title:** EVALUATING PROTEIN SIGNATURES

(57) **Abstract:** Test arrays of capture probes are used to identify characteristic information about a sample. For example, the methods can be used to identify the presence of a cancer cell or a pathogen in a sample from a subject, or the presence of a target molecule in an environmental sample.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US03/17979

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : C12Q 1/68; C07H 21/02, 21/04
US CL : 435/6; 536/23.1, 24.3

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 435/6; 536/23.1, 24.3

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
WEST

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	WO 01/40803 A1 (DIVERSYS LTD) 7 June 2001, see entire document.	1-3, 19,24,25,36,37,40,41 ----------- 4-18,20-23,26-35,42-49

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
13 October 2005 (13.10.2005)	04 NOV 2005

Name and mailing address of the ISA/US	Authorized officer
Mail Stop PCT, Attn: ISA/US Commissioner of Patents P.O. Box 1450 Alexandria, Virginia 22313-1450 Facsimile No. (571) 273-3201	Valerie Bell-Harris for Teresa E. Strzelecka Telephone No. (571) 272-1600